

Genome-wide analysis of individual coding variants and HLA-II-associated self-immunopeptidomes in ulcerative colitis – Supplementary Material

Supplementary Figures:

Supplementary Figure 1: Quantile-quantile plot of association summary statistics of the imputed genotyping data.....	3
Supplementary Figure 2: Manhattan plot of the imputed genotyping data.	4
Supplementary Figure 3: Regional association plot for 1p36.22 with the top hit chr1:12090328 (rs72641067).	5
Supplementary Figure 4: Regional association plot for the locus 1p36.22 with the top hit chr1:12601409(rs12136952).	6
Supplementary Figure 5: Regional association plot for the locus 1p36.13 with the top hit chr1:19839478 (rs7523442).	7
Supplementary Figure 6: Regional association plot for the locus 2p16.1 with the top hit chr2:59899466 (rs17050481).	8
Supplementary Figure 7: Regional association plot for the locus 3q28 with the top hit chr3:189167658 (rs73184427).	9
Supplementary Figure 8: Regional association plot for the locus 4p12 with the top hit chr4:45120035 (rs113429955).	10
Supplementary Figure 9: Regional association plot for the locus 5p14.3 with the top hit chr5:18748431 (rs2937516).	11
Supplementary Figure 10: Regional association plot for the locus 5p13.1 with the top hit chr5:40323836 (rs348594).	12
Supplementary Figure 11: Regional association plot for the locus 6p21.33 with the top hit chr6:31118325 (rs117198148).	13
Supplementary Figure 12: Regional association plot for the locus 6p21.32 with the top hit chr6:32644620 (rs6927022).	14
Supplementary Figure 13: Regional association plot for the locus 9q22.2 with the top hit chr9:89844860 (rs36147380).	15
Supplementary Figure 14: Regional association plot for the locus 10q24.2 with the top hit chr10:99541336 (rs4590800).	16
Supplementary Figure 15: Regional association plot for the locus 12p13.31 with the top hit chr12:9333053 (rs187033004).	17
Supplementary Figure 16: Regional association plot for the locus 12q24.13 with the top hit chr12:113880288 (rs3782449).	18
Supplementary Figure 17: Regional association plot for the locus 16q12.1 with the top hit chr16:50666737 (rs139397276).	19
Supplementary Figure 18: Regional association plot for the locus 16q22.1 with the top hit chr16:67493201 (rs77919558).	20
Supplementary Figure 19: Regional association plot for the locus 19q13.11 with the top hit chr19:32088213 (rs6510221).	21
Supplementary Figure 20: Regional association plot for the locus 19q13.31 with the top hit chr19:43804850 (rs364691).	22
Supplementary Figure 21: Regional association plot for the locus 22q13.1 with the top hit chr22:39318699 (rs1569498).	23

Supplementary Figure 22: Quantile-quantile plot of association summary statistics of the whole exome data.	23
Supplementary Figure 23: Manhattan plot of the exome data.	24
Supplementary Figure 24: Regional association plot for the locus 1p36.13 in the exome data with the top hit chr1:19890366 (rs7523442).	25
Supplementary Figure 25: Regional association plot for the locus 6p21.32 in the exome data with the top hit chr6:32661551 (rs28724240).	26
Supplementary Figure 26: Regional association plot for the locus 12p13.2 in the exome data with the top hit chr12:11092079 (rs113197337).	27
Supplementary Figure 27: Regional association plot for the locus 12q24.33 in the exome data with the top hit chr12:132711135 (rs7973452).	28
Supplementary Figure 28: Regional association plot for the locus 22q11.21 in the exome data with the top hit chr22:20429371 (rs755163625).	29
Supplementary Figure 29: Associations at the NOD2 locus and the influence on the protein level.	31
Supplementary Figure 30: Power analysis based on the GWAS catalog data.	31
Supplementary Figure 31: Finemapping of the HLA region.	32
Supplementary Figure 32: Dendrogram of HLA-DRB1 alleles.	32
Supplementary Figure 33: Binding logo plot of associated HLA-DR alleles in differentiation to alleles with the other direction of effect.	33
Supplementary Figure 34: Binding logo plot of associated HLA-DQ alleles.	34
Supplementary Figure 35: Vulcano plot of the PepWAS analysis.	35

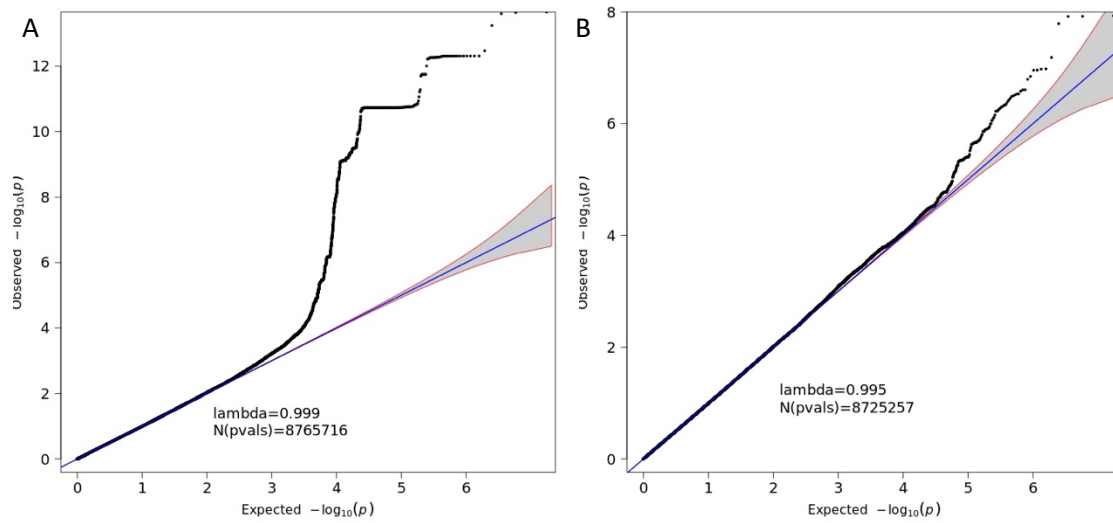
Supplementary Tables:

Supplementary Table 1: Sample number before, during, and after QC.	36
Supplementary Table 2: Genes and transcripts used to generate the proteome.	36
Supplementary Table 3: At least nominal significantly associated lead variants identified in the “imputed genotyping” dataset or the “Exome” dataset.	36
Supplementary Table 4: The association results with the HLA imputed data.	37
Supplementary Table 5: The significant associated peptides from the PepWAS analysis.	37

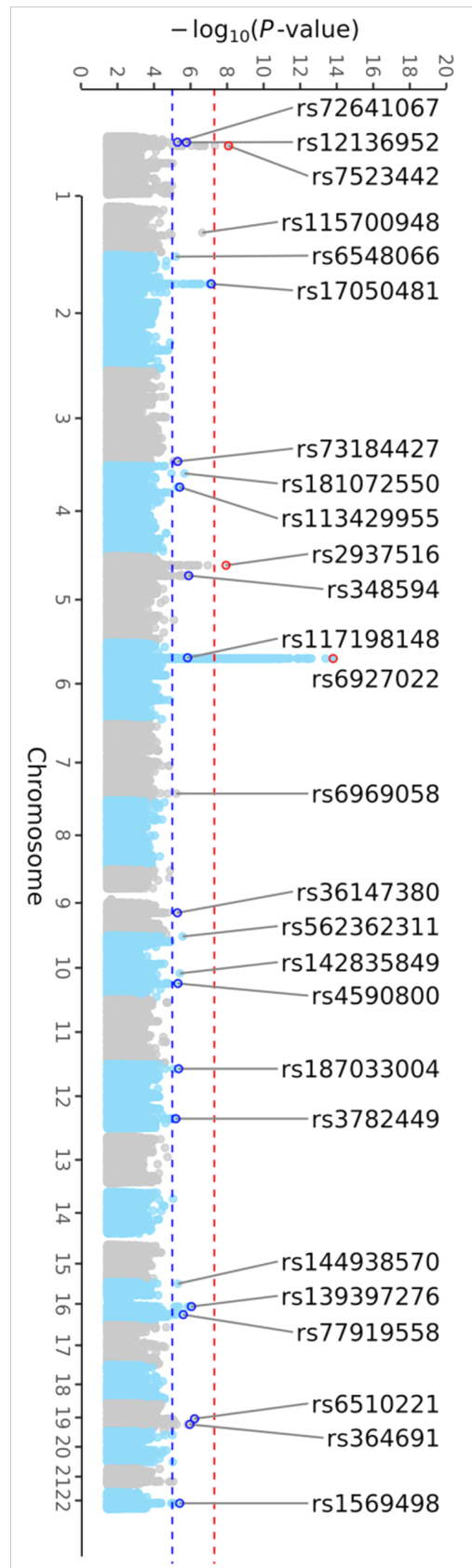
Supplementary References

References:	38
--------------------------	----

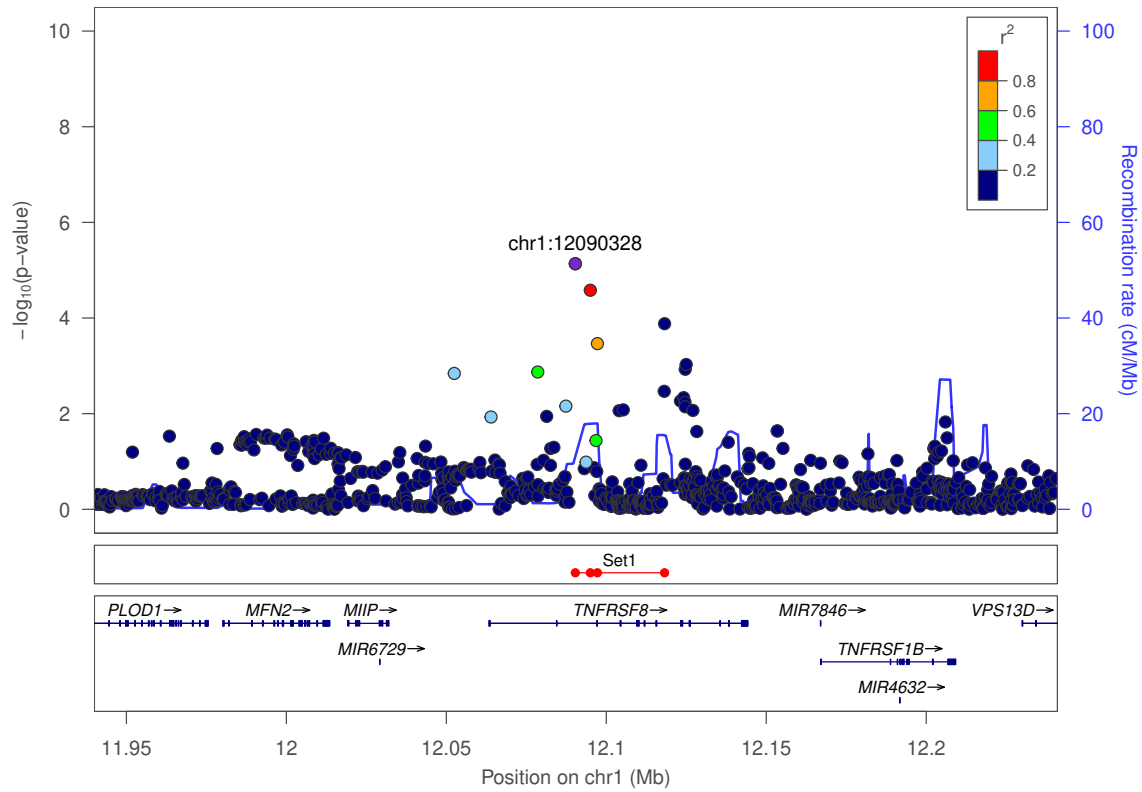
Supplementary Figures



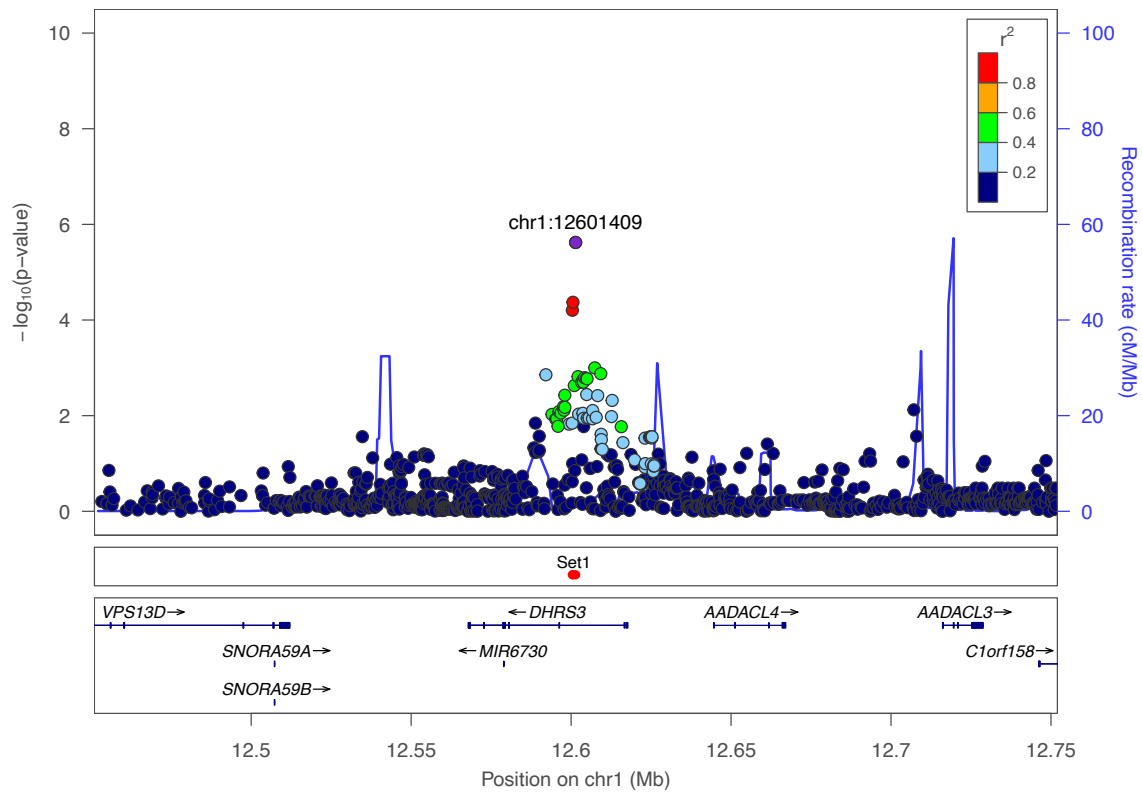
Supplementary Figure 1: Quantile-quantile plot of association summary statistics of the imputed genotyping data. The 95% concentration band under random sampling is shown in gray. The genomic inflation factor λ is defined as the ratio of the medians of the sample χ^2 test statistics and the 1-df χ^2 distribution (0.455).¹ Panel (A) includes all 8,765,716 variants with MAF > 1% and an imputation score $r^2 > 0.6$. Panel (B) excludes the variants of the HLA-region (chr6:29-34MB).



Supplementary Figure 2: Manhattan plot of the imputed genotyping data with a MAF > 1% and an imputation score $r^2 > 0.6$. All loci of at least nominal significance (blue horizontal line; $P < 1 \times 10^{-5}$) are annotated by the SNP-ID. Loci with LD support are highlighted with a blue (nominal significance) or red circle (genome-wide significance, red horizontal line; $P < 5 \times 10^{-8}$).

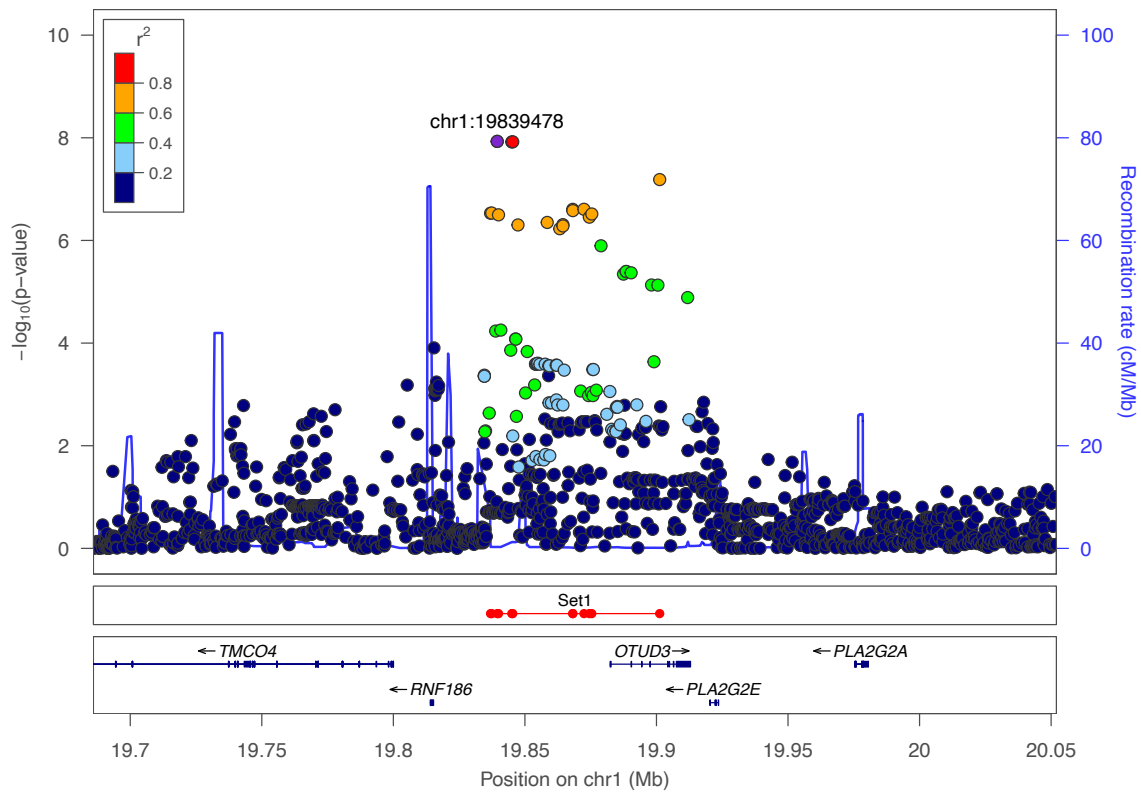


Supplementary Figure 3: Regional association plot for 1p36.22 with the top hit chr1:12090328 (rs72641067). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom².

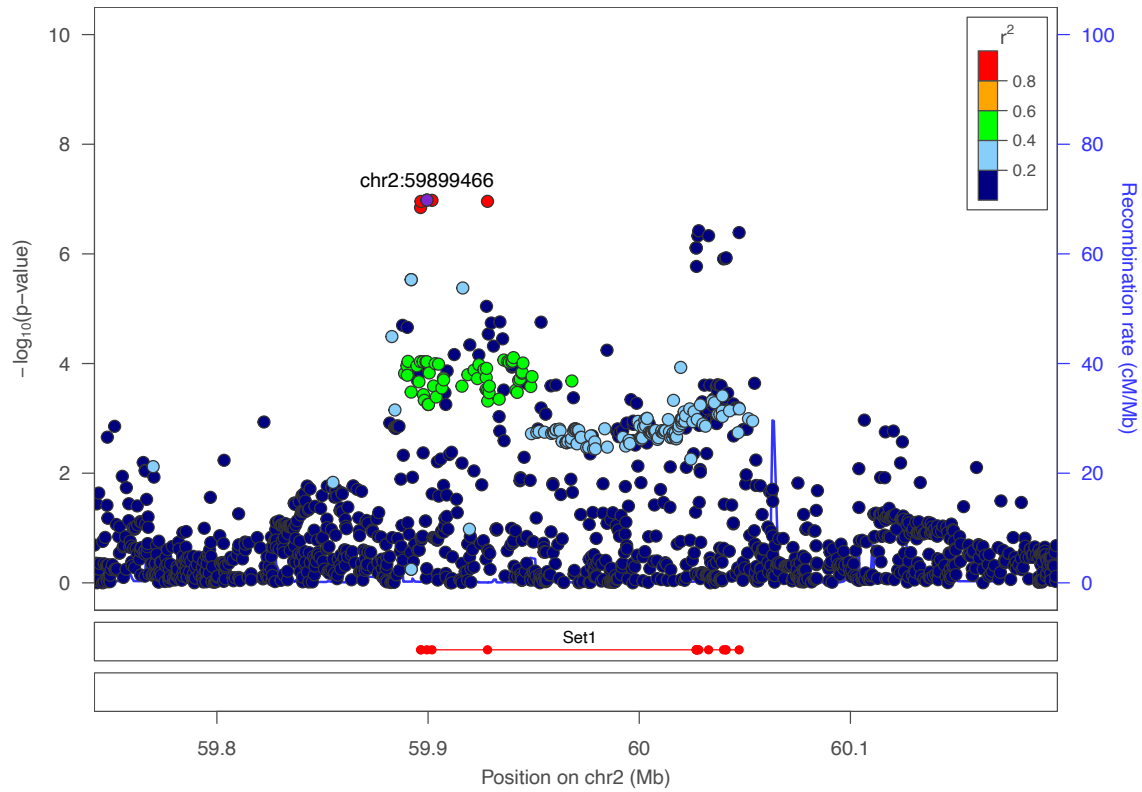


Supplementary Figure 4: Regional association plot for the locus 1p36.22 with the top hit chr1:12601409(rs12136952). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the

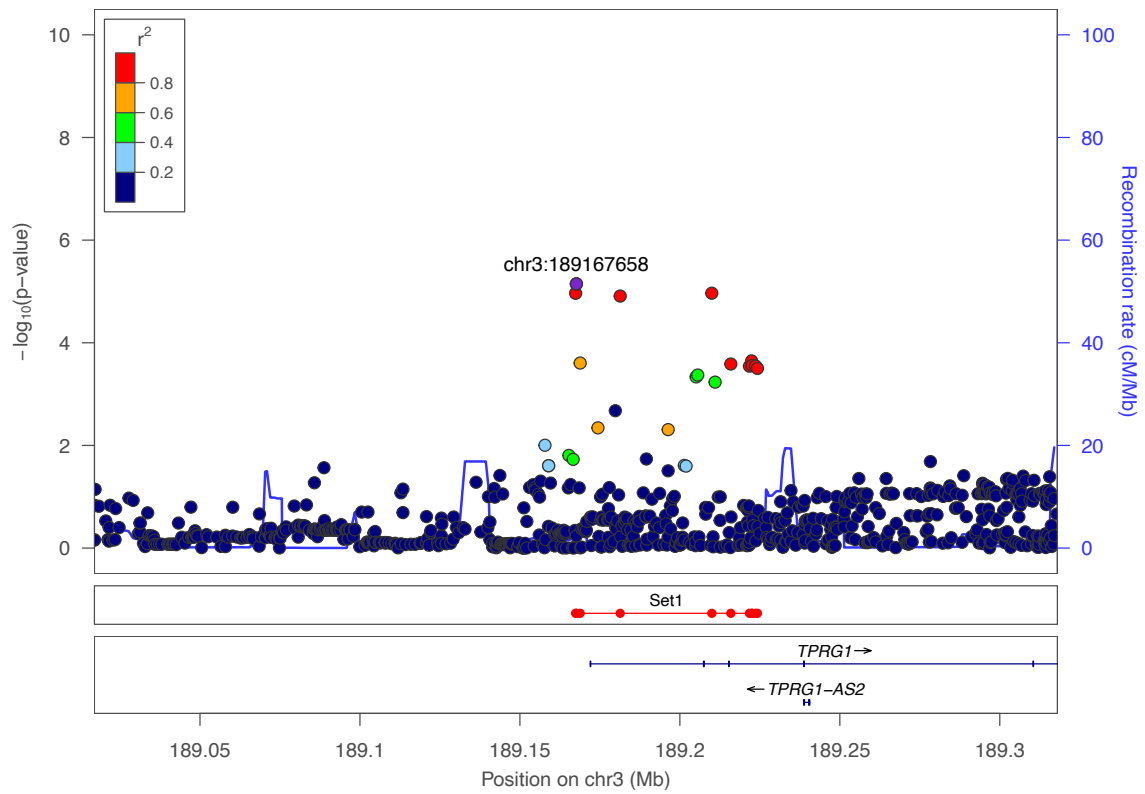
gene. The plot was created using LocusZoom².



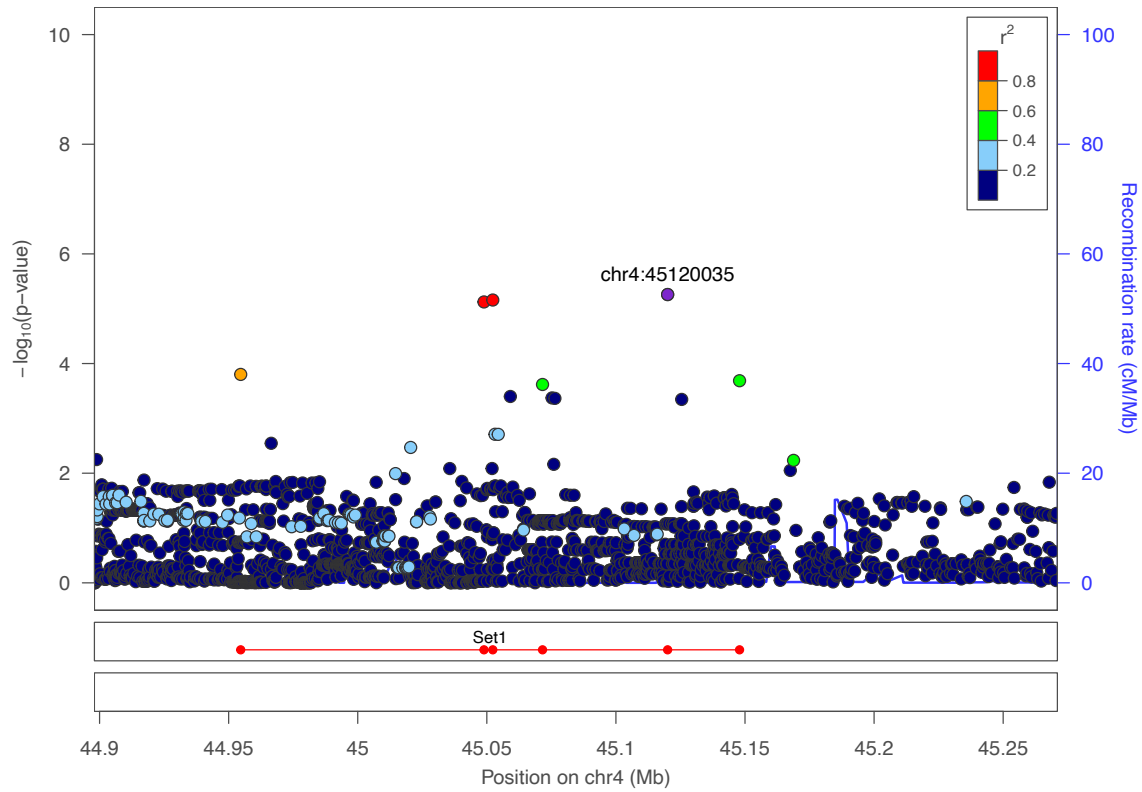
Supplementary Figure 5: Regional association plot for the locus 1p36.13 with the top hit chr1:19839478 (rs7523442). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom².



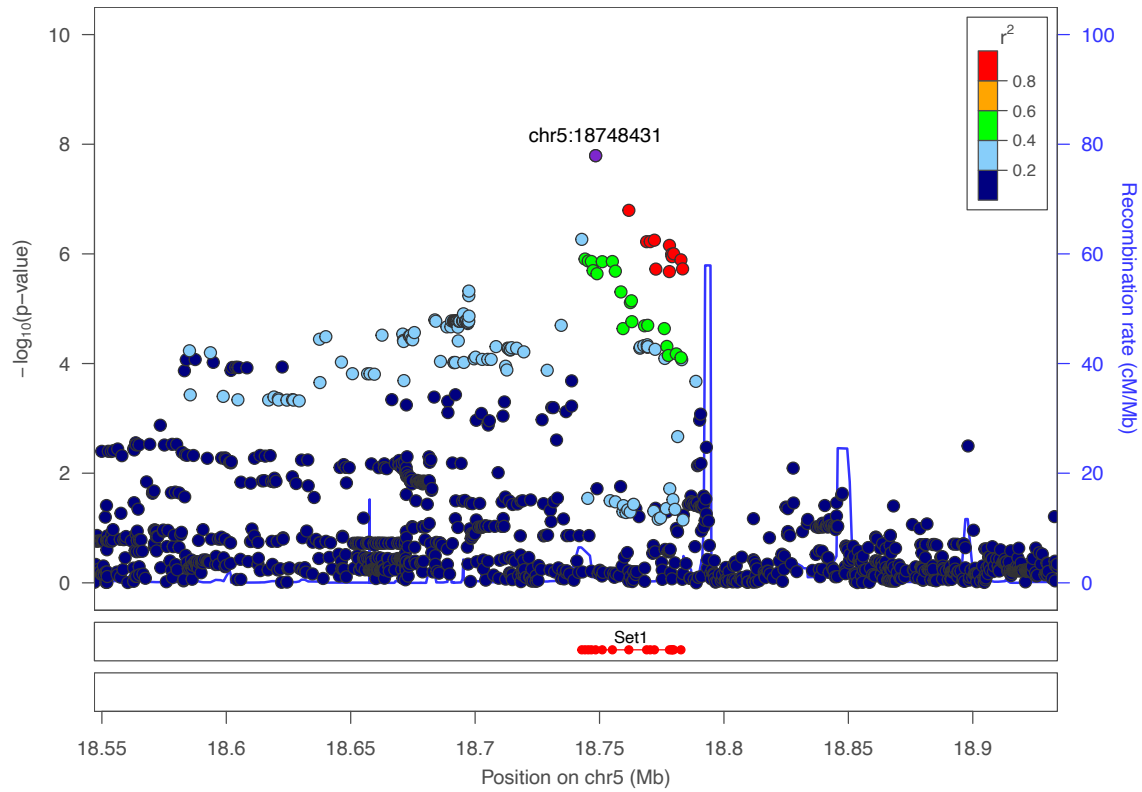
Supplementary Figure 6: Regional association plot for the locus 2p16.1 with the top hit chr2:59899466 (rs17050481). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom².



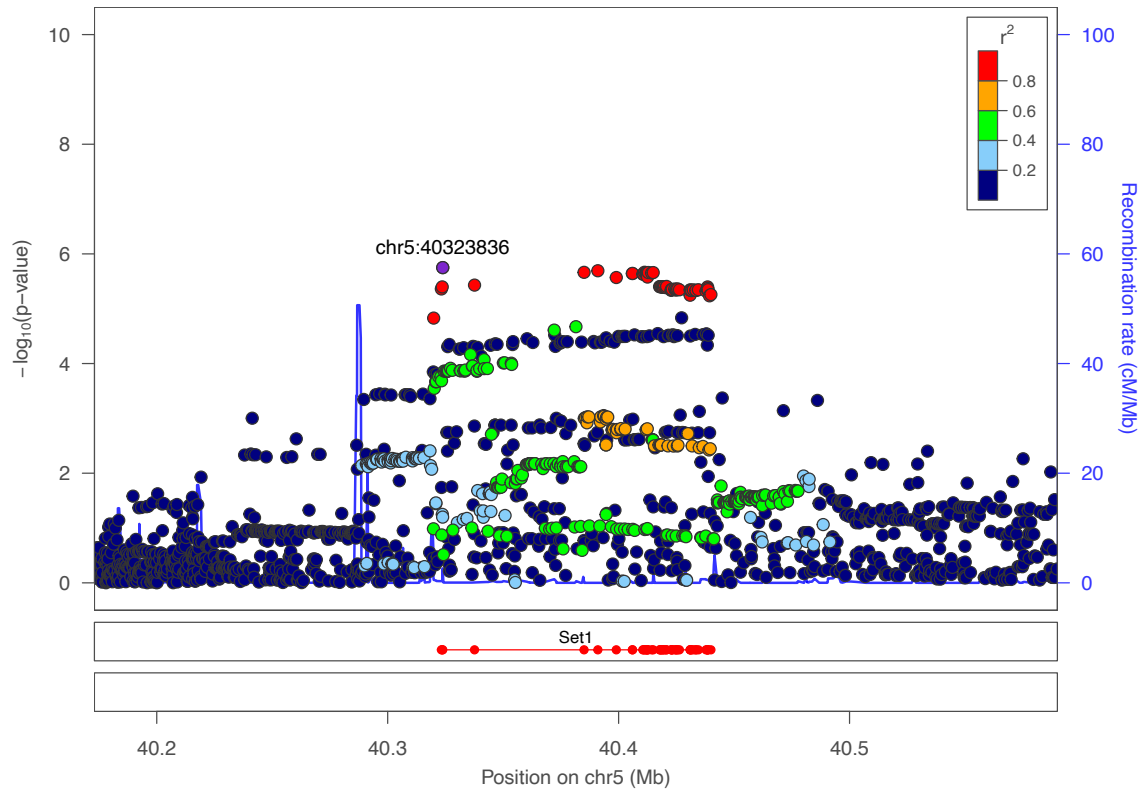
Supplementary Figure 7: Regional association plot for the locus 3q28 with the top hit chr3:189167658 (rs73184427). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom².



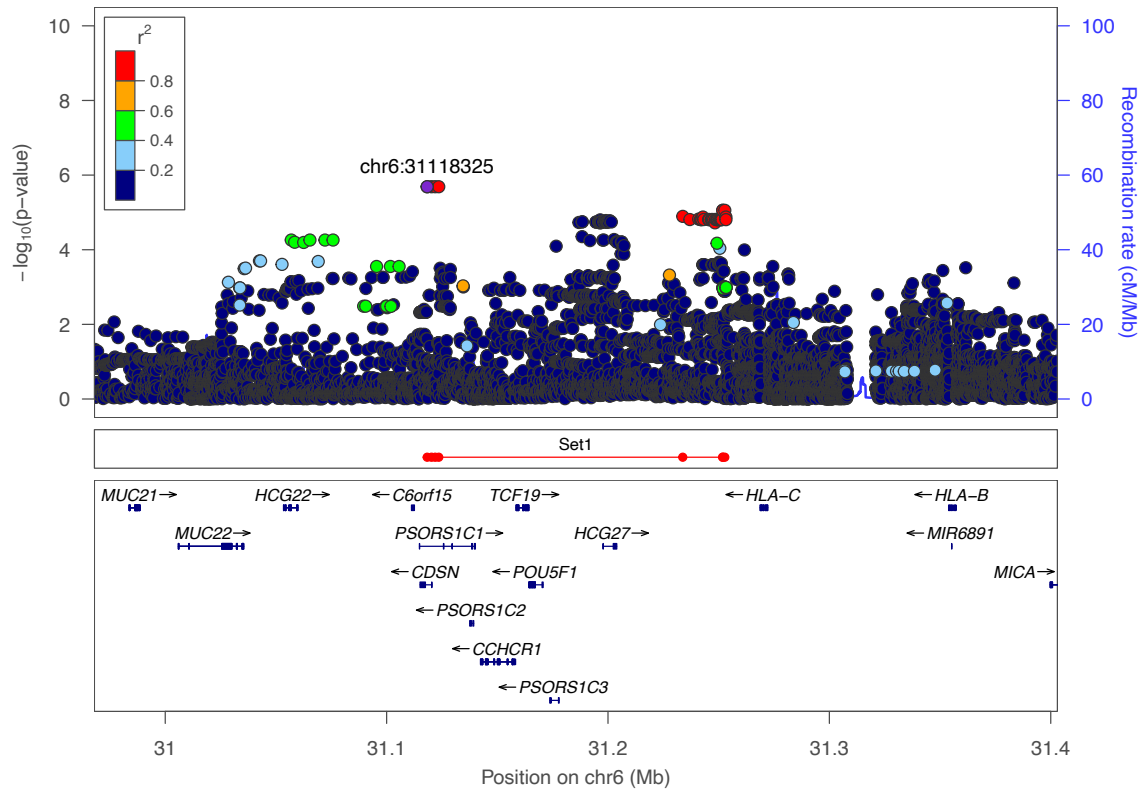
Supplementary Figure 8: Regional association plot for the locus 4p12 with the top hit chr4:45120035 (rs113429955). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom².



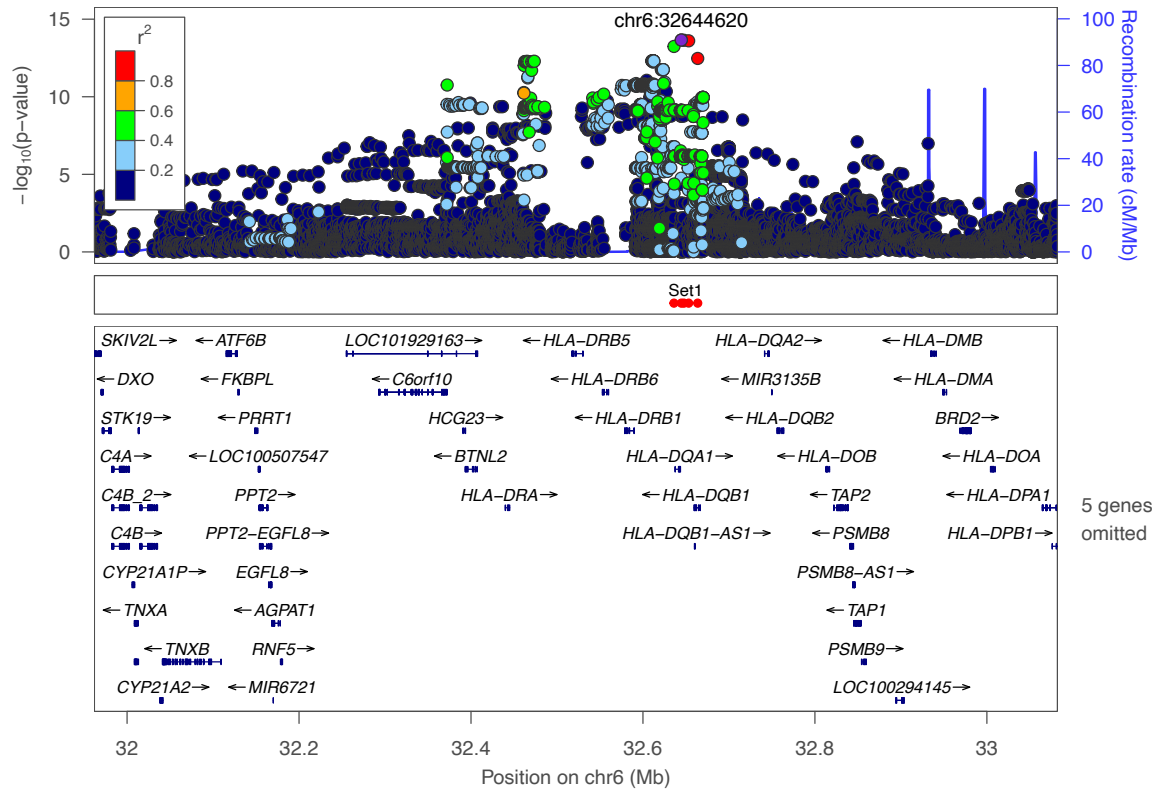
Supplementary Figure 9: Regional association plot for the locus 5p14.3 with the top hit chr5:18748431 (rs2937516). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom².



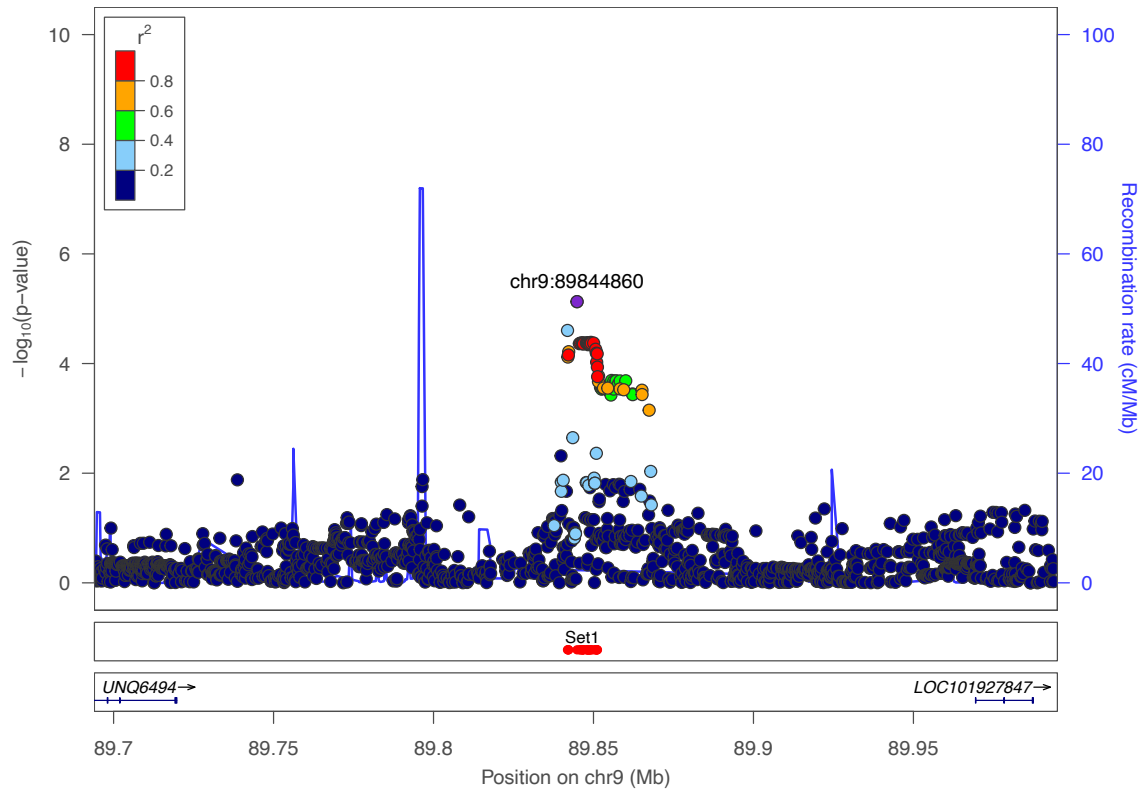
Supplementary Figure 10: Regional association plot for the locus 5p13.1 with the top hit chr5:40323836 (rs348594). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom².



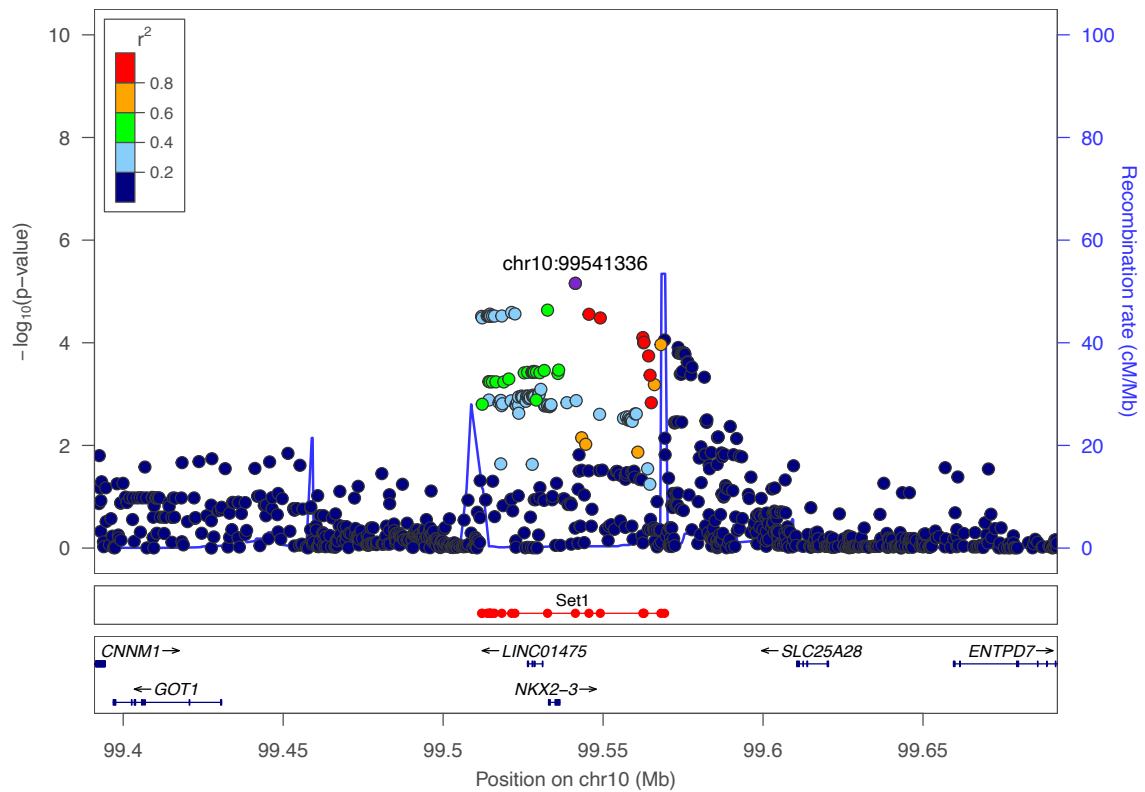
Supplementary Figure 11: Regional association plot for the locus 6p21.33 with the top hit chr6:31118325 (rs117198148). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom².



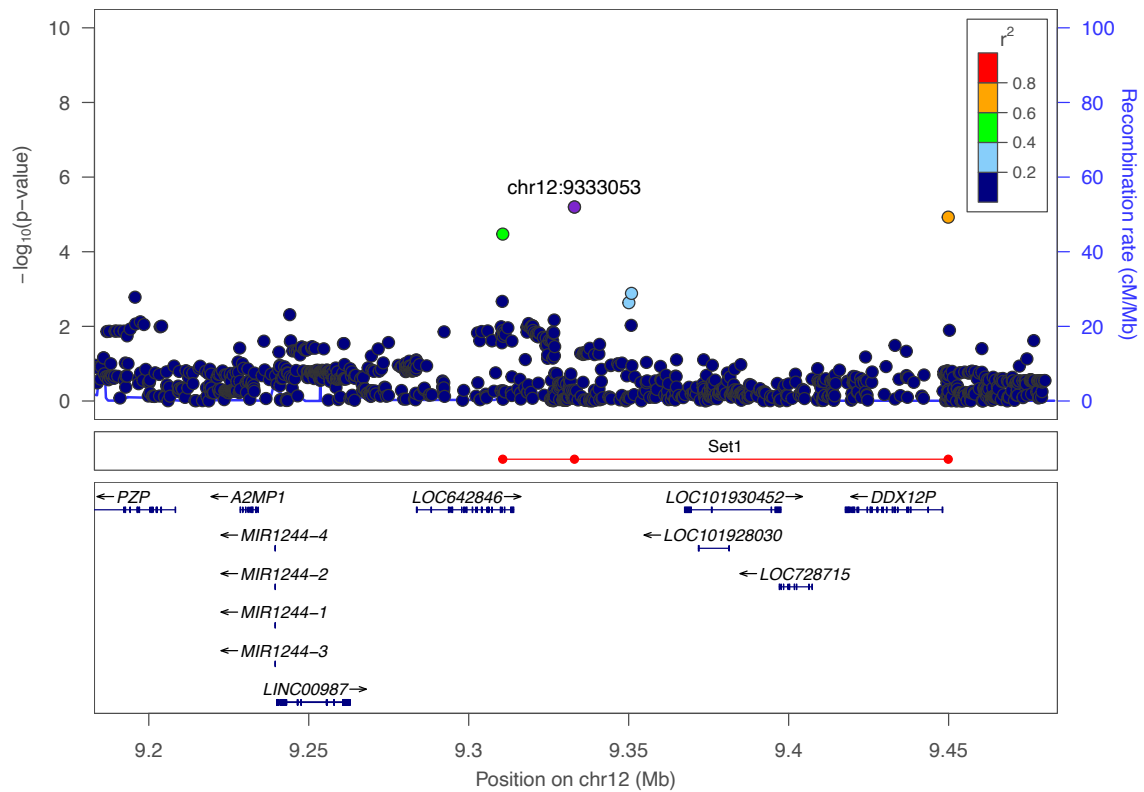
Supplementary Figure 12: Regional association plot for the locus 6p21.32 with the top hit chr6:32644620 (rs6927022). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom².



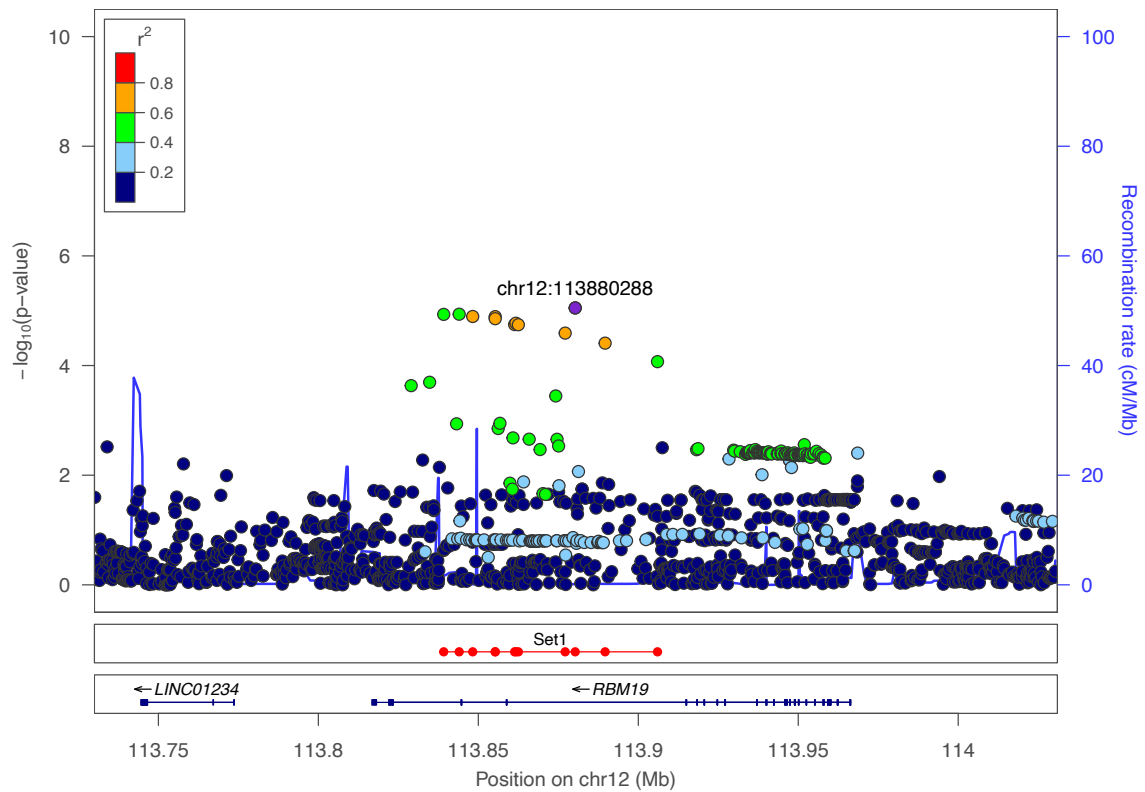
Supplementary Figure 13: Regional association plot for the locus 9q22.2 with the top hit chr9:89844860 (rs36147380). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom².



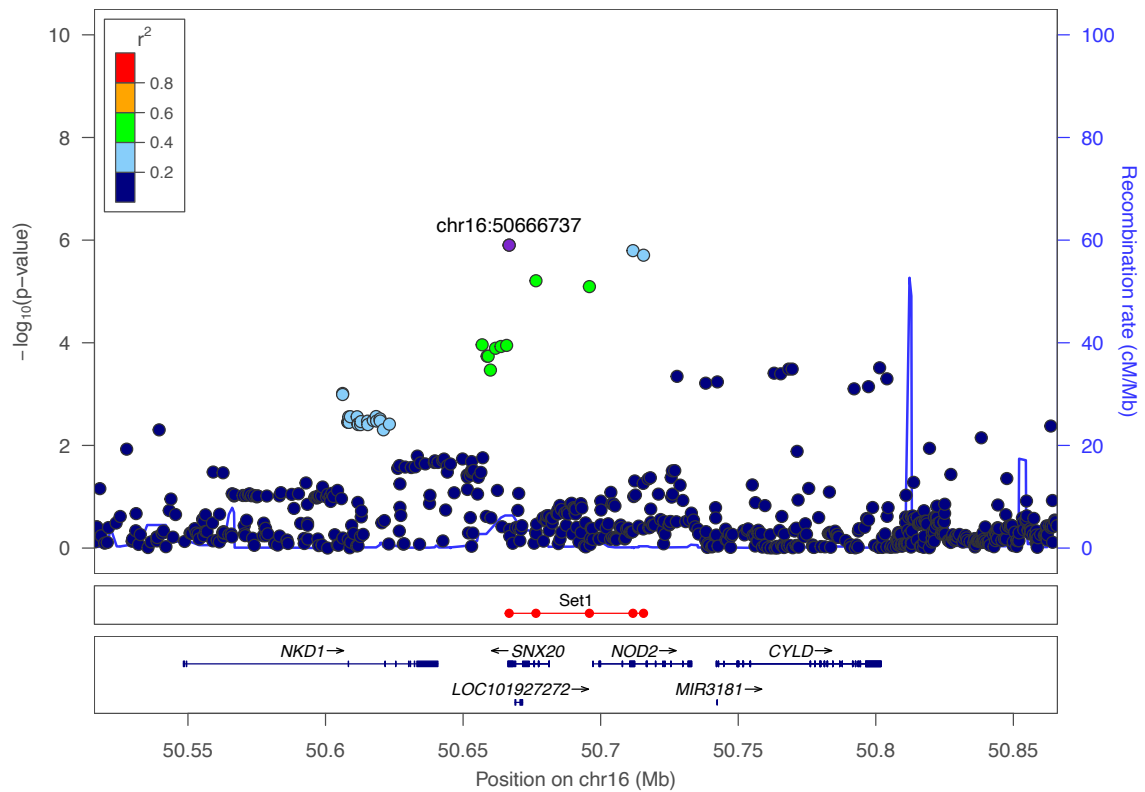
Supplementary Figure 14: Regional association plot for the locus 10q24.2 with the top hit chr10:99541336 (rs4590800). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom².



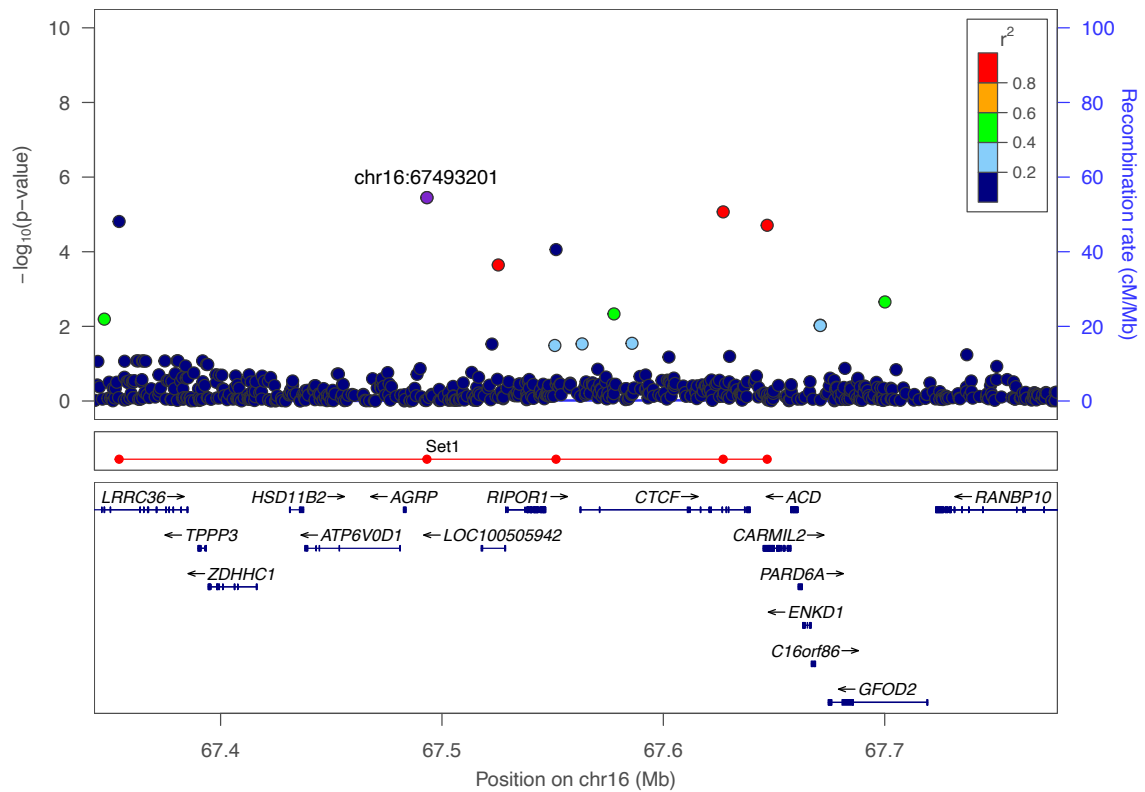
Supplementary Figure 15: Regional association plot for the locus 12p13.31 with the top hit chr12:9333053 (rs187033004). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom².



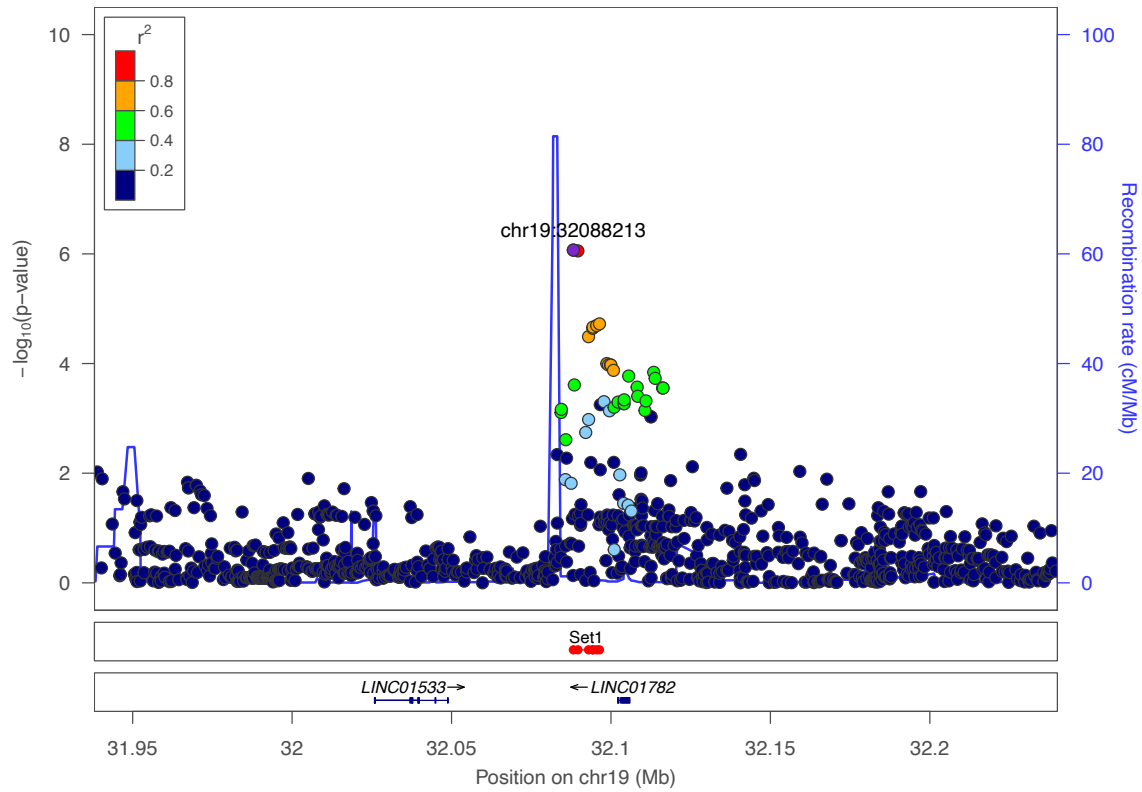
Supplementary Figure 16: Regional association plot for the locus 12q24.13 with the top hit chr12:113880288 (rs3782449). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom².



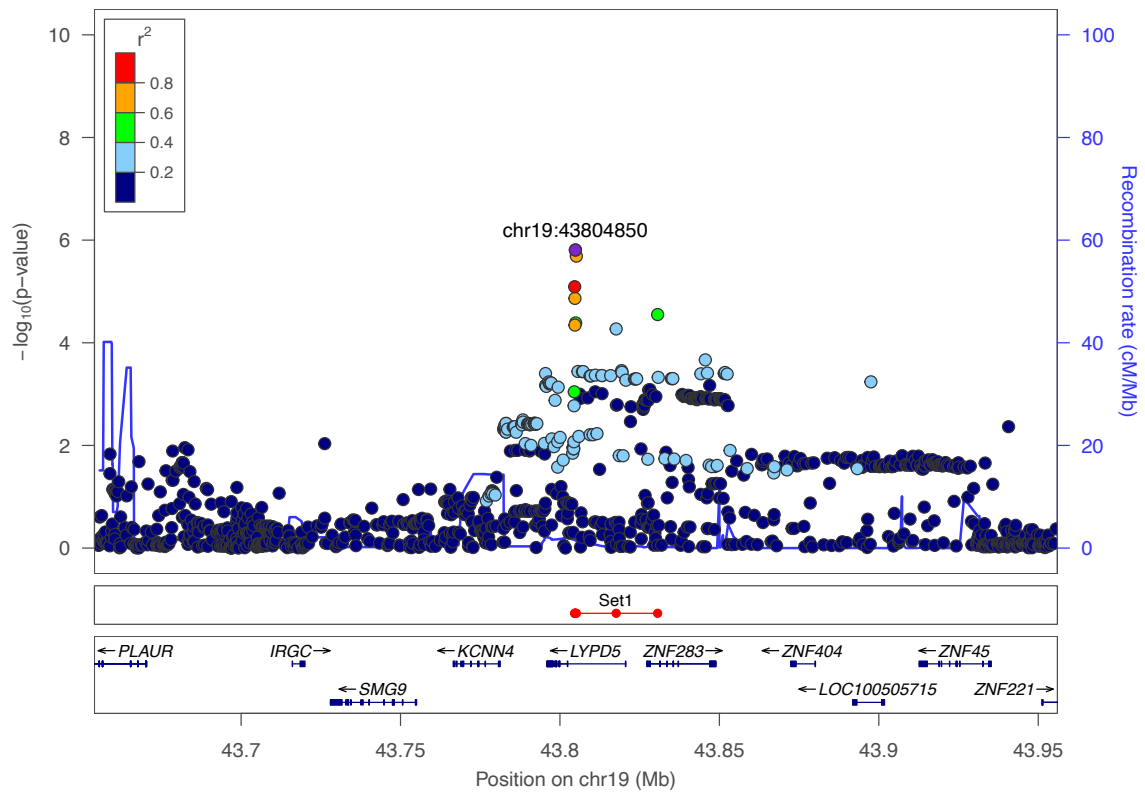
Supplementary Figure 17: Regional association plot for the locus 16q12.1 with the top hit chr16:50666737 (rs139397276). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom².



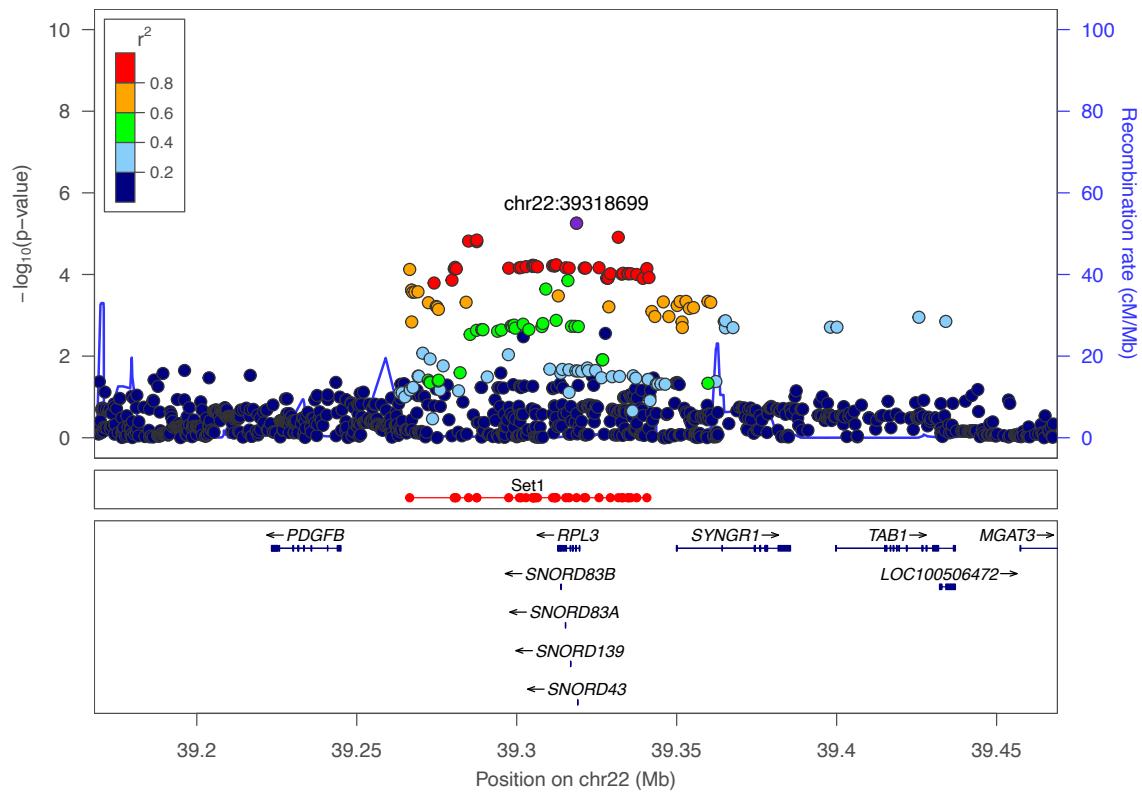
Supplementary Figure 18: Regional association plot for the locus 16q22.1 with the top hit chr16:67493201 (rs77919558). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom².



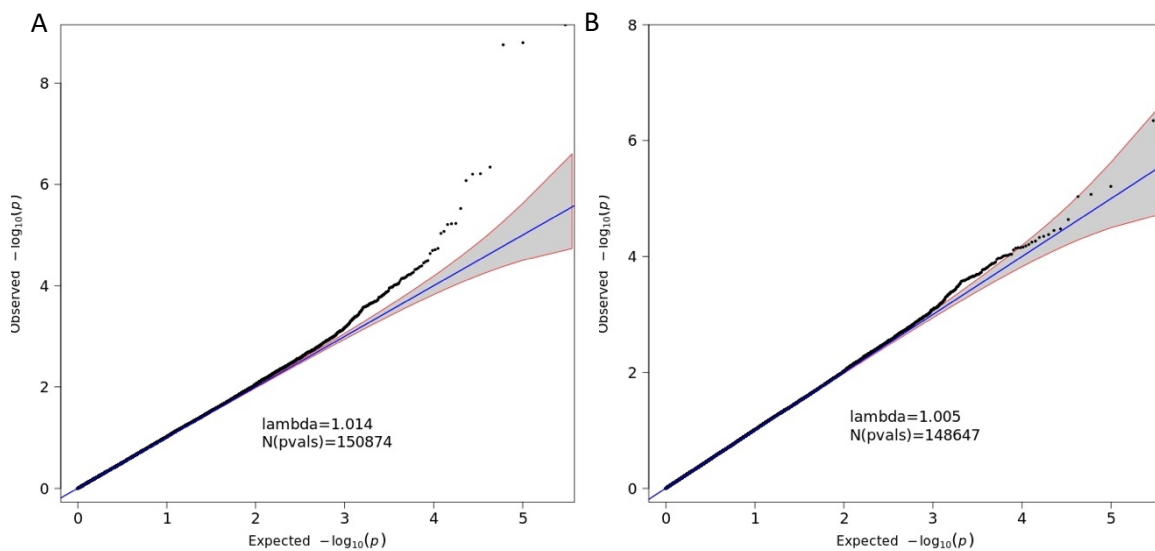
Supplementary Figure 19: Regional association plot for the locus 19q13.11 with the top hit chr19:32088213 (rs6510221). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom².



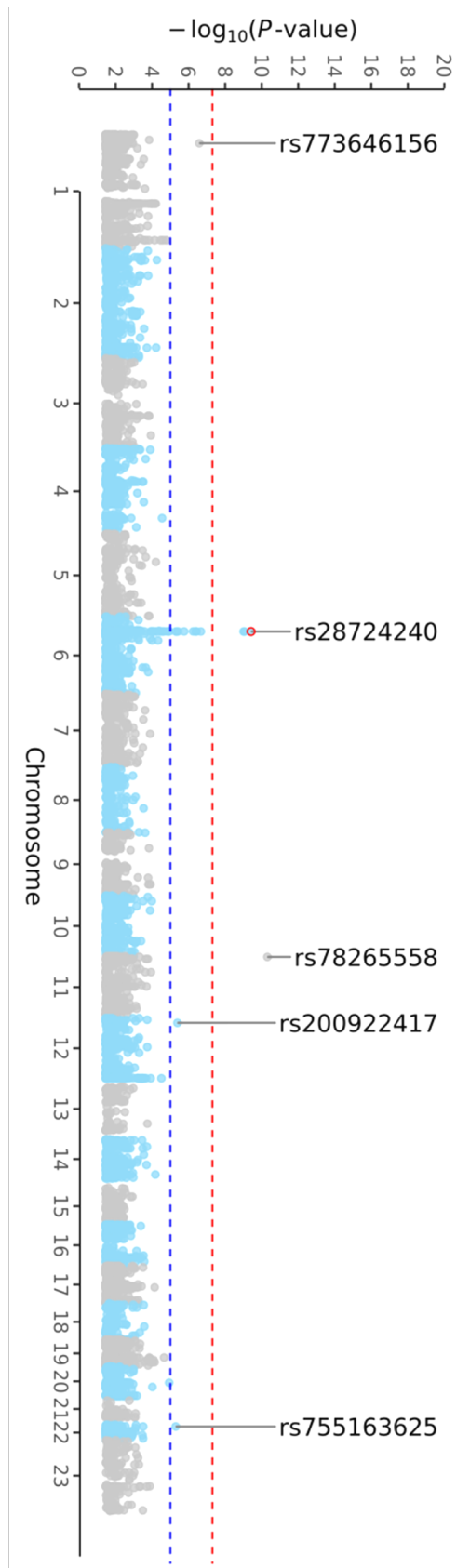
Supplementary Figure 20: Regional association plot for the locus 19q13.31 with the top hit chr19:43804850 (rs364691). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom².



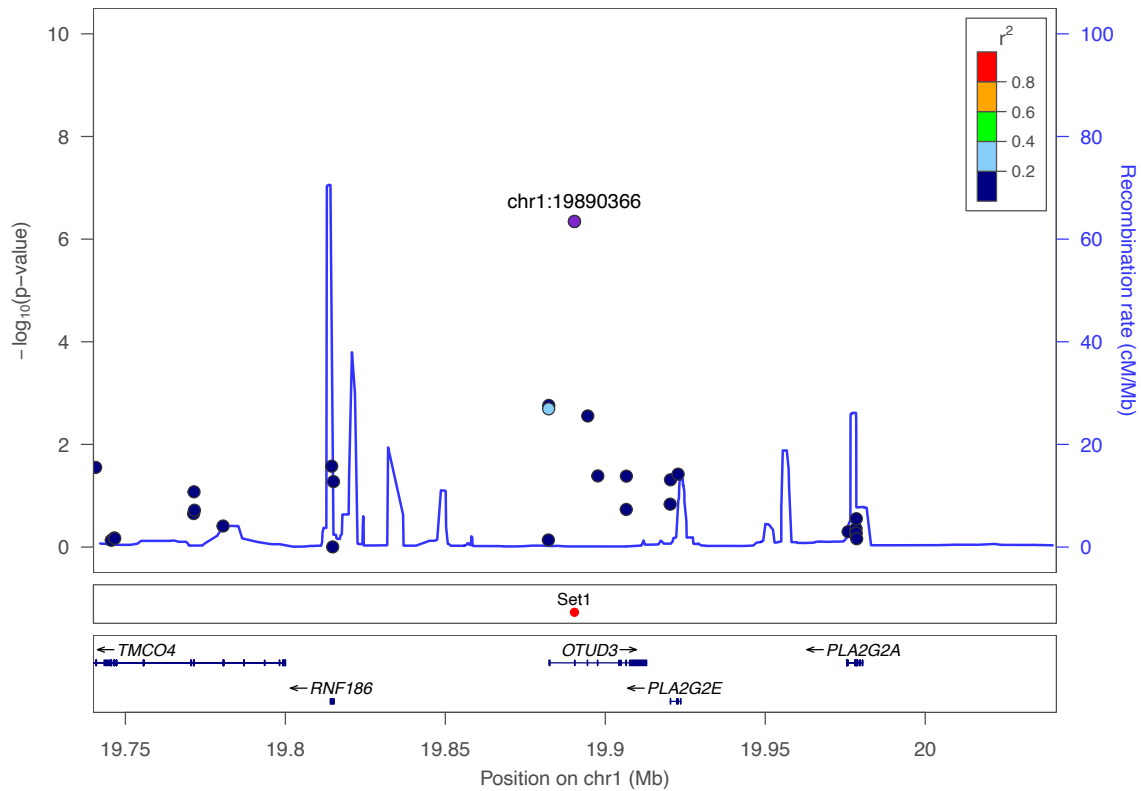
Supplementary Figure 21: Regional association plot for the locus 22q13.1 with the top hit chr22:39318699 (rs1569498). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom².



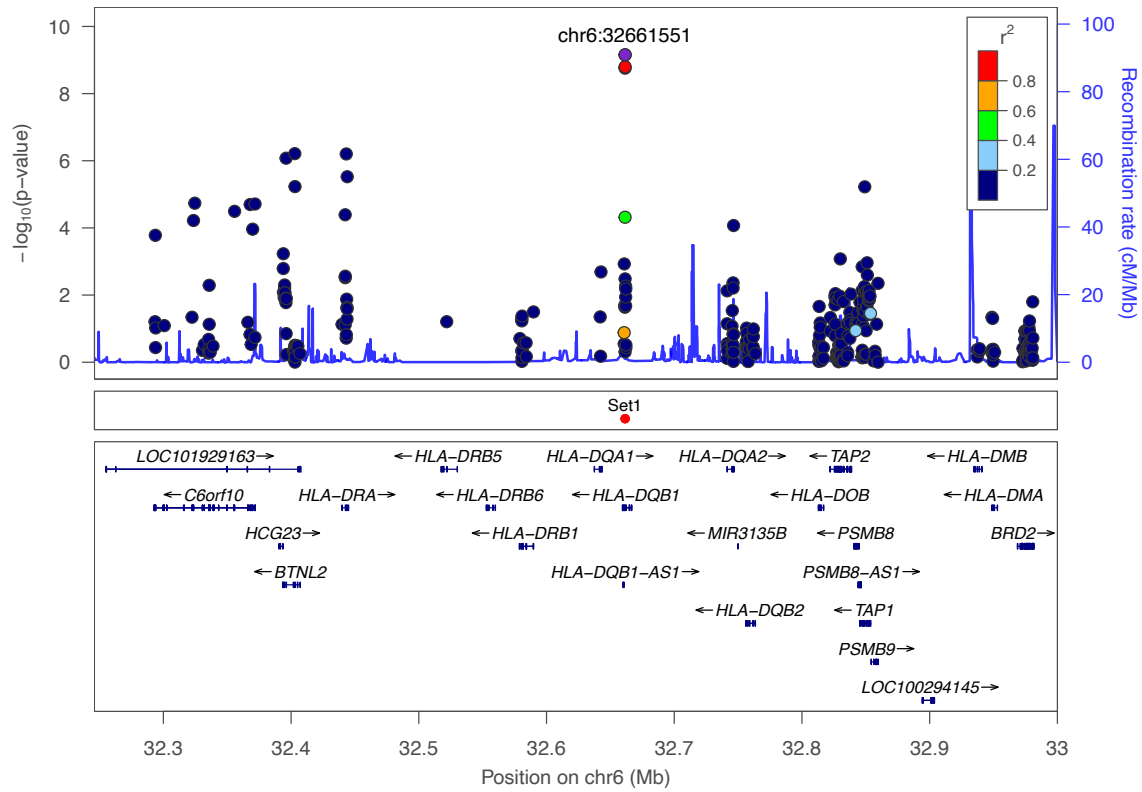
Supplementary Figure 22: Quantile-quantile plot of association summary statistics of the whole exome data. The 95% concentration band under random sampling is shown in gray. The genomic inflation factor lambda is defined as the ratio of the medians of the sample χ^2 test statistics and the 1-df χ^2 distribution (0.455).¹ The left figure includes all 150,874 variants with MAF > 1% and an imputation score $r^2 > 0.6$. The right figure excludes the variants of the HLA-region (chr6:29-34MB).



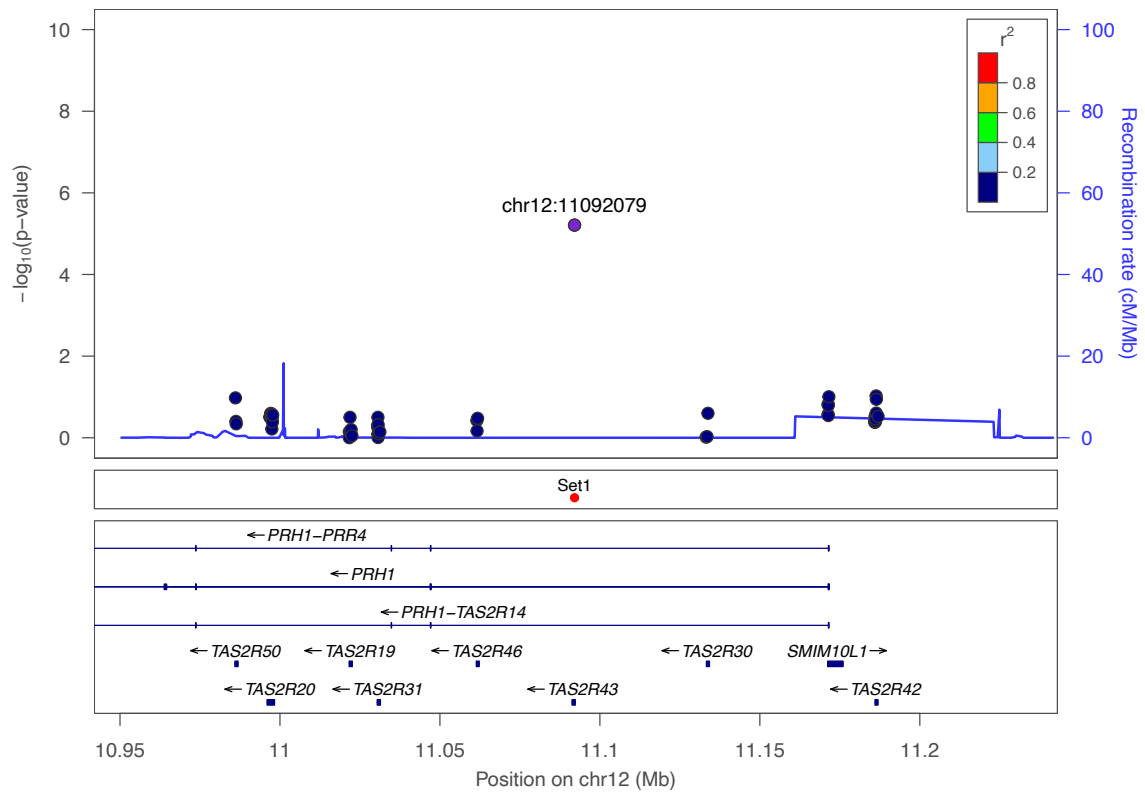
Supplementary Figure 23: Manhattan plot of the exome data with a MAF >1% and an imputation score $r^2 > 0.6$. All loci of at least nominal significance (blue horizontal line; $P < 1 \times 10^{-5}$) are annotated by the SNP-ID. Loci with LD support are highlighted with a blue (nominal significance) or red circle (genome-wide significance, red horizontal line; $P < 5 \times 10^{-8}$).



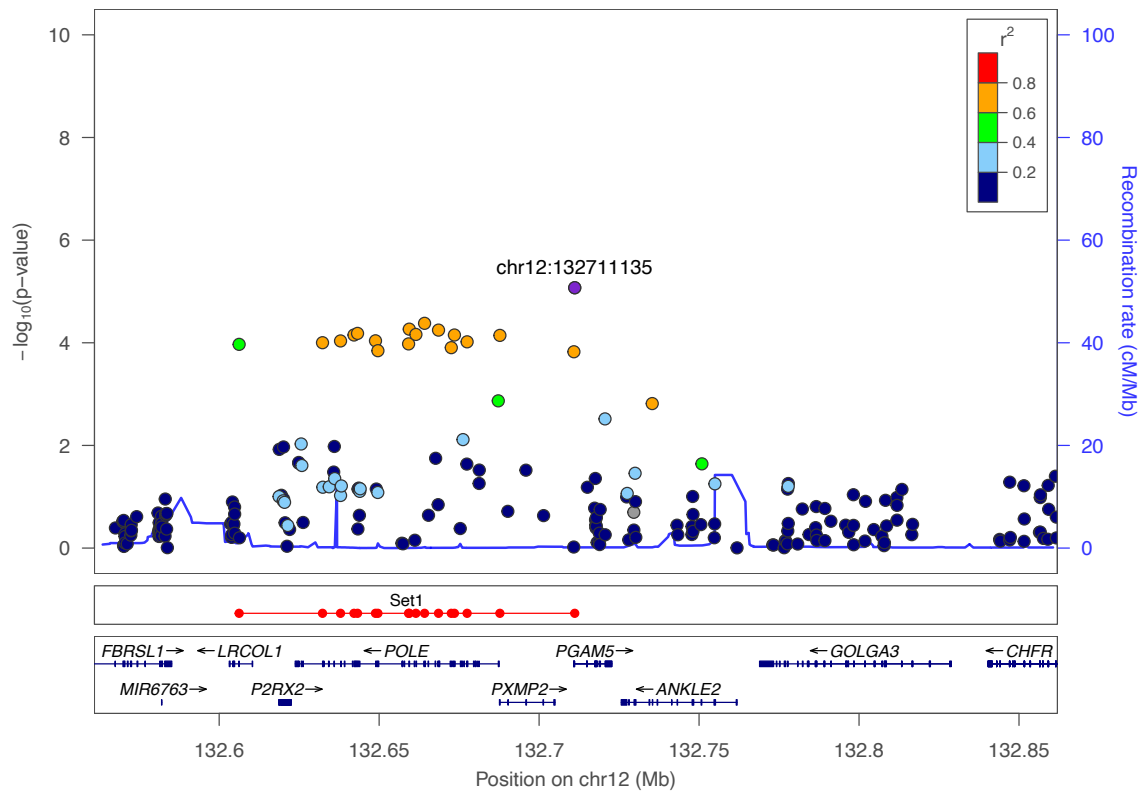
Supplementary Figure 24: Regional association plot for the locus 1p36.13 in the exome data with the top hit chr1:19890366 (*rs7523442*). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom².



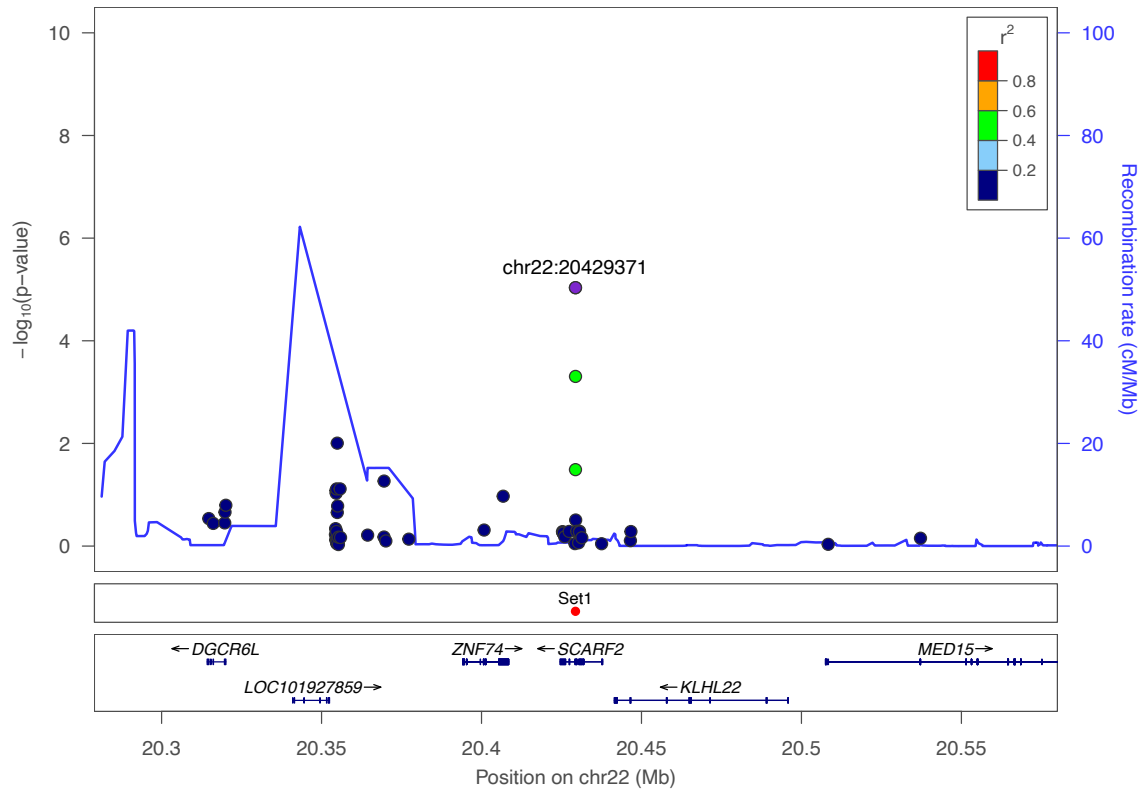
Supplementary Figure 25: Regional association plot for the locus 6p21.32 in the exome data with the top hit chr6:32661551 (rs28724240). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom².



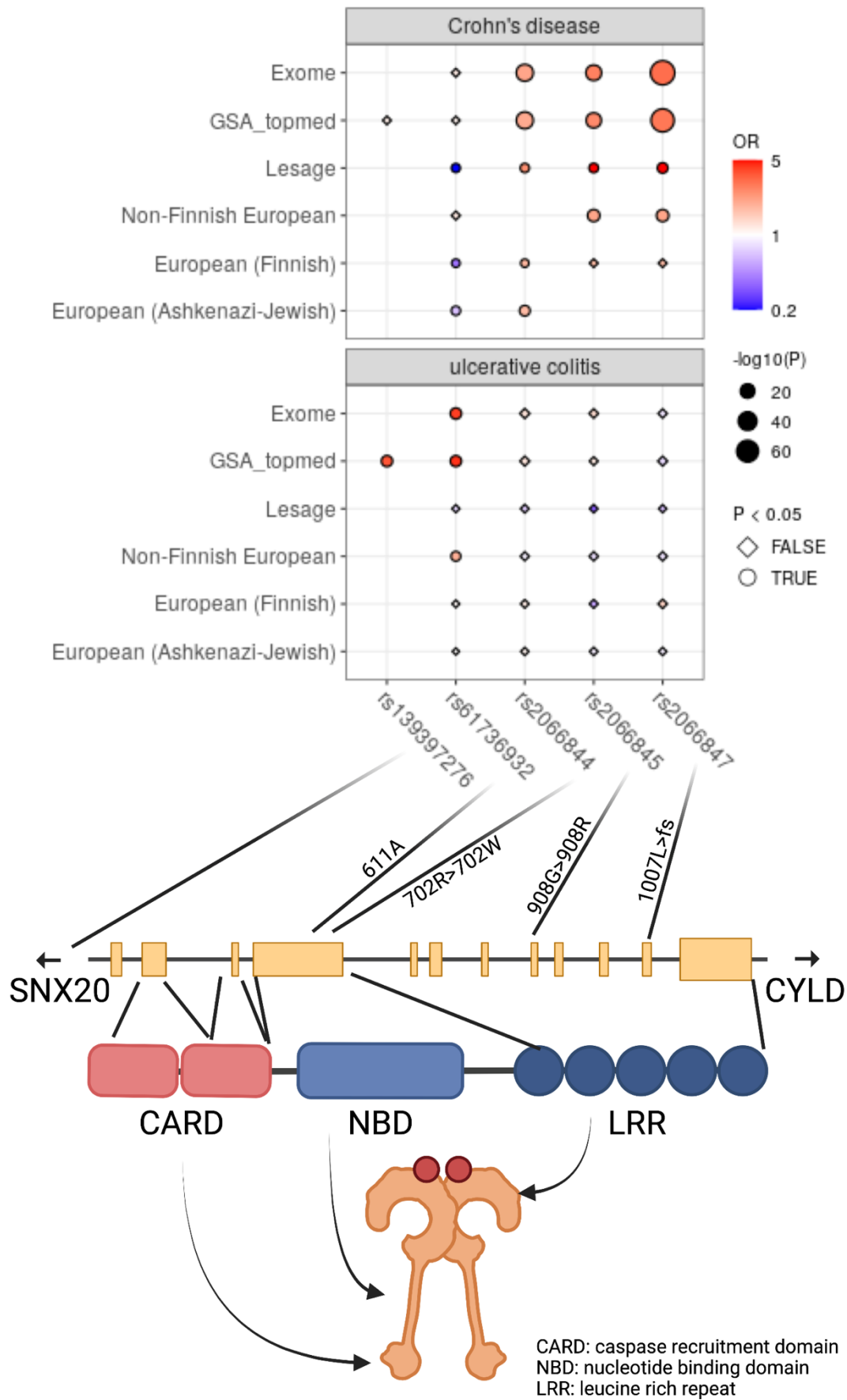
Supplementary Figure 26: Regional association plot for the locus 12p13.2 in the exome data with the top hit chr12:11092079 (rs113197337). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom².



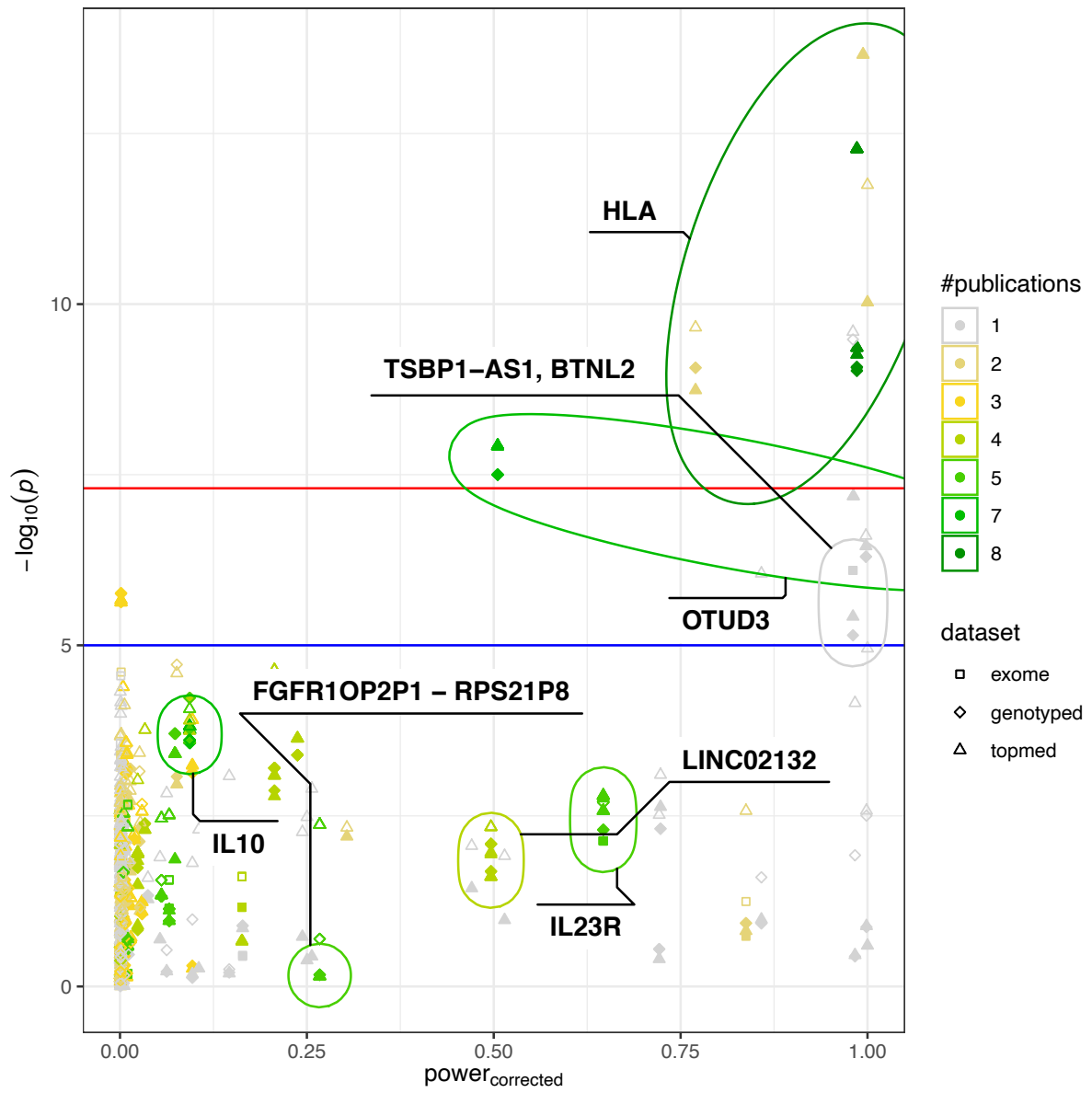
Supplementary Figure 27: Regional association plot for the locus 12q24.33 in the exome data with the top hit *chr12:132711135* (*rs7973452*). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom².



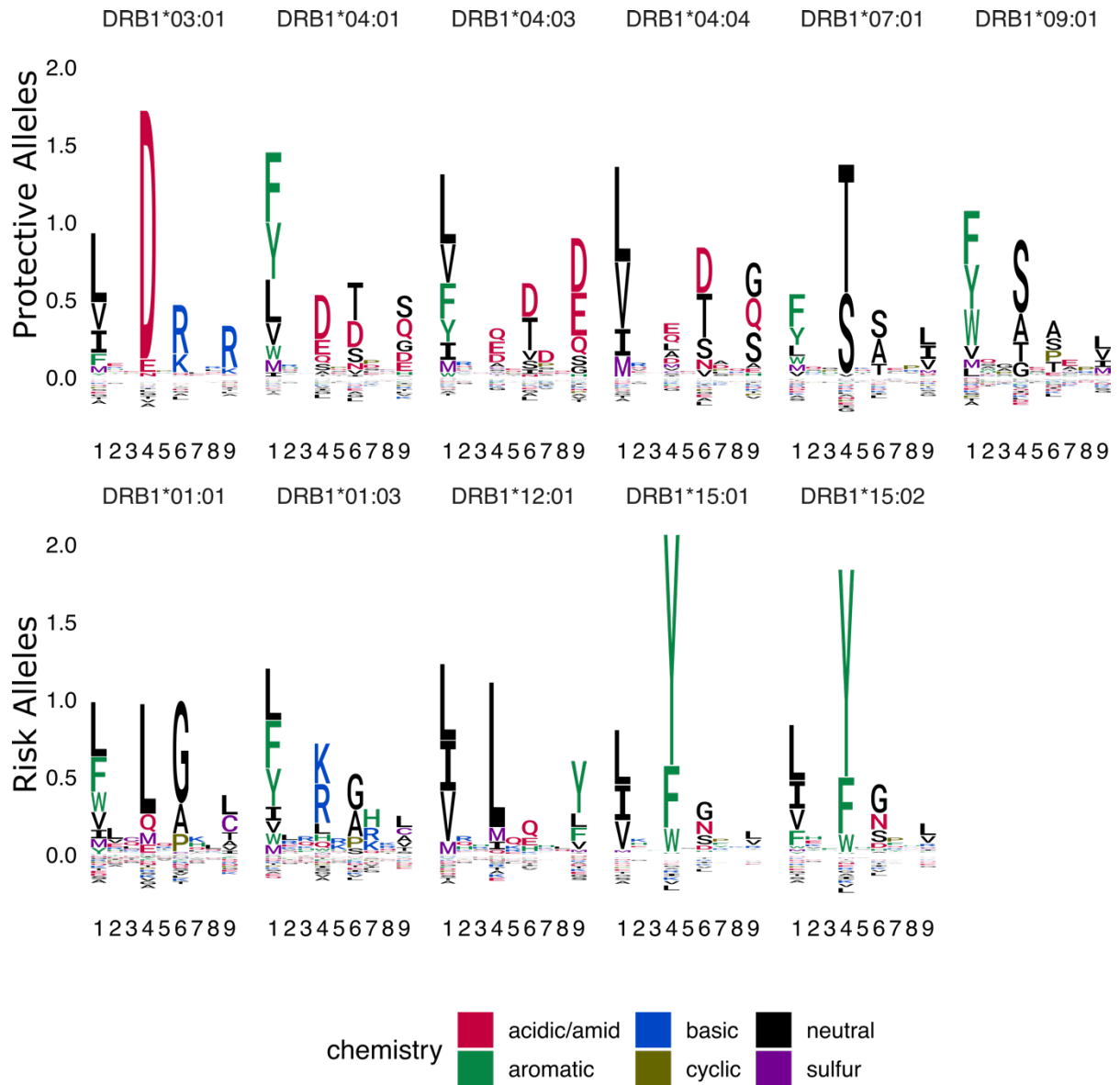
Supplementary Figure 28: Regional association plot for the locus 22q11.21 in the exome data with the top hit *chr22:20429371* (*rs755163625*). The purple dot represents the most strongly associated SNP with ulcerative colitis. The color of the dots represents the linkage disequilibrium (LD) with the most strongly associated SNP (see color legend). The positions represent the genome build GRCh38. The recombination rate is shown in centimorgans (cM) per million base pairs (Mb). The bottom part shows the name and locations of the genes within the region. The thicker blue line represents the position of the exons, while the thinner line represents the intronic regions. The direction of transcription is represented by an arrow behind the name of the gene. The plot was created using LocusZoom².



Supplementary Figure 29: Associations at the *NOD2* locus and the influence on the protein level. *Rs139397277* is our main signal and *rs61736932* the strongest associated variant within the exome data while the other three variants are those previously identified as associated with CD. Created with BiorRender.com.

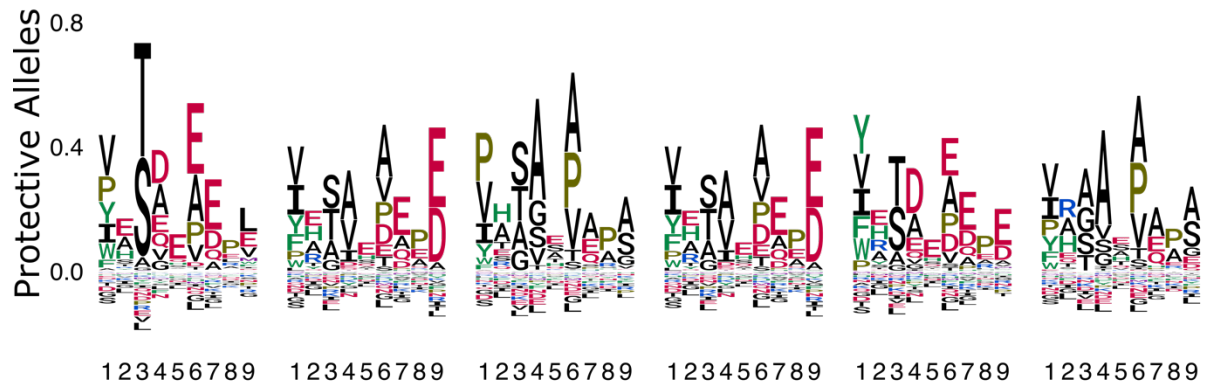


Supplementary Figure 30: Power analysis based on the GWAS catalog data. The corrected power is calculated based on the median odds ratios and standard errors as listed in the GWAS catalog, when removing the strongest effect if more than one association is given. The power is calculated for the nominal significance level 1×10^{-5} and the frequencies given in our data.



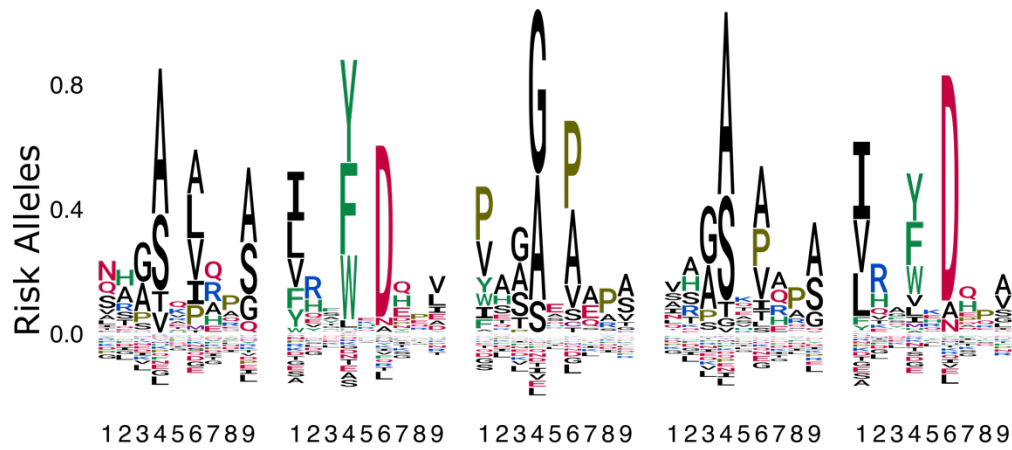
Supplementary Figure 33: Binding logo plot of associated HLA-DR alleles in differentiation to alleles with the other direction of effect. The upper row represents the protective associated alleles, the bottom line the logos of the risk alleles. The motifs are based on the NetMHCIIpan-4.0 predictions of the binding cores of all peptides at least annotated as weak binders, excluding peptides binding against one of the alleles of different direction of effect. The single letters represent the one letter amino acid code colored by the chemical properties of the amino acids.

DQ*02:01|02:02 DQ*03:01|03:02 DQ*02:01|03:03 DQ*03:03|03:02 DQ*03:03|02:02 DQ*03:03|03:03



1 2 3 4 5 6 7 8 9 1 2 3 4 5 6 7 8 9 1 2 3 4 5 6 7 8 9 1 2 3 4 5 6 7 8 9 1 2 3 4 5 6 7 8 9

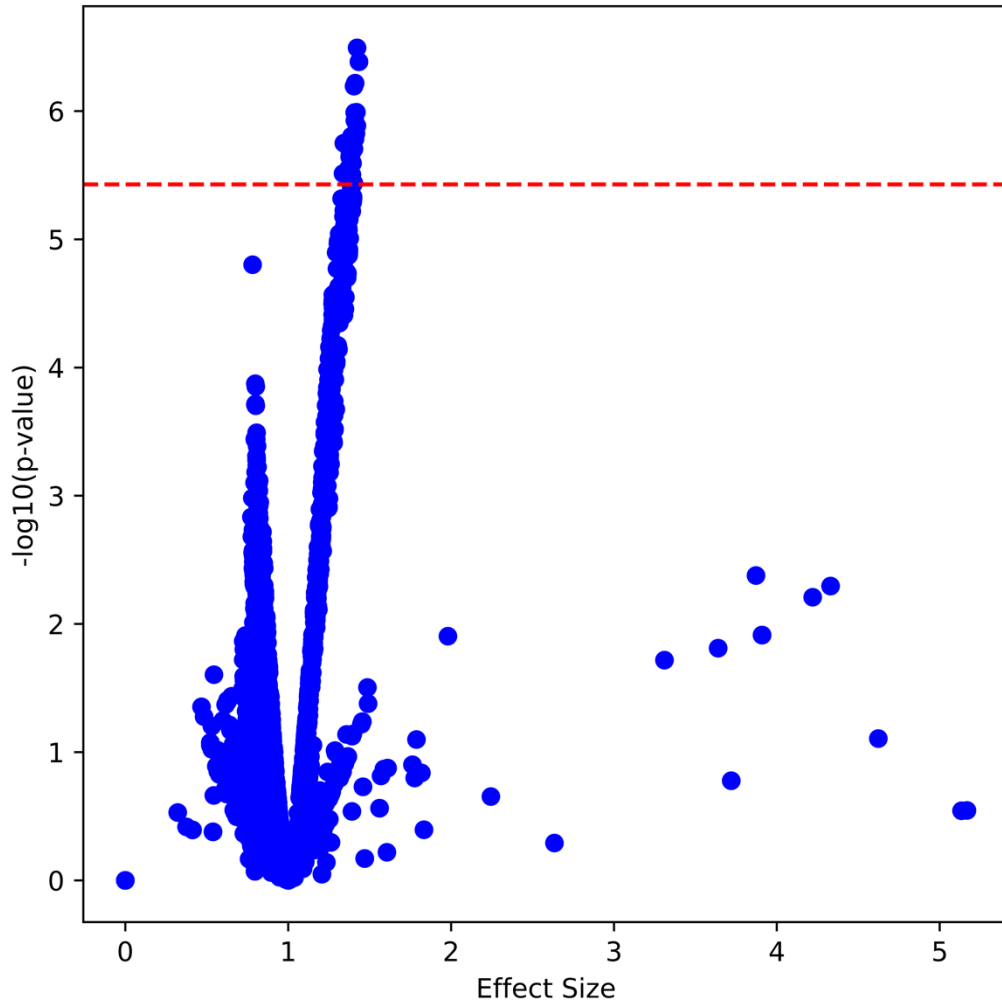
DQ*01:02|01:06:02 DQ*01:01|01:05:01 DQ*05:05|03:01 DQ*01:02|01:06:01 DQ*01:02|01:05:01



1 2 3 4 5 6 7 8 9 1 2 3 4 5 6 7 8 9 1 2 3 4 5 6 7 8 9 1 2 3 4 5 6 7 8 9

chemistry ■ acidic/amid ■ basic ■ neutral
■ aromatic ■ cyclic ■ sulfur

Supplementary Figure 34: Binding logo plot of associated HLA-DQ alleles. The allele names are shortened, e.g., DQA1*02:01-DQB1*02:02 is written as DQ*02:01|02:02. The motifs are based on the NetMHCIIpan-4.0 predictions of the binding cores of all peptides at least annotated as weak binders. The single letters represent the one letter amino acid code colored by the chemical properties of the amino acids.



Supplementary Figure 35: *Vulcano plot of the PepWAS analysis. Each dot represents the PepWAS results for one peptide. The red dotted line represents the Bonferroni corrected P-value ($P\text{-value} < 3.73 \times 10^{-6}$ based on 13,411 peptides).*

Supplementary Tables

Supplementary Table 1: Sample number before, during, and after QC. Numbers in white lines represent sample numbers to be removed.

	UC	Control	All traits
Metadata	950	4681	20563
GSA+Exome (GSA/Exome)			19769 (20554/19772)
GSA input	950	4680	20554
-non German samples (Metadata)	0	0	1312
-only Exome data	0	0	3
-only GSA data	31	171	785
-duplicate samples	1	46	256
-unique blacklist	32	217	2065
GSA QC input			18492
-Missingness outlier	12	77	392
-Heterozygosity outlier	2	5	41
-PCA outliers			506
-duplicates	0	0	2
-unique QC removed	53	121	903
GSA after QC	865	4342	17589
-Relatives	2	157	1672
-relatives not already removed			1634
GSA final Qced	863	4185	15955
Exome input	119	4509	19772
-het/hom	3	11	43
-TiTv	0	0	1
-singletons	6	11	65
-missingness	0	0	7
-sex	0	0	13
-unique QC removed	6	16	84
Exome final Qced	913	4493	19688
GSA and Exome Qced (including relatives)			17138
Association (no relatives, only UC and Controls)	863	4185	5048

Supplementary Table 2: Genes and transcripts used to generate the proteome. Given are next to the genetic hg38 location and the ensemble-ids also the biotype of the transcript, the hgnc symbol and the uniprot gene ids.

[See separate xlsx-file]

Supplementary Table 3: At least nominal significantly associated lead variants identified in the “imputed genotyping” dataset or the “Exome” dataset. For each locus an identifier is given in the “NR” column and the band information, further in case of a locus top hit it is annotated whether the locus is LD supported. For comparison additional entries from external sources were added. Those sources are the GWAS catalog (dataset named by the first author of the publication; namely: Liu JZ³, de Lange K⁴, Silverberg MS⁵, McGovern DP⁶, Jostins L⁷, Barrett JC⁸, Anderson CA⁹, Asano K¹⁰, Okamoto D¹¹, Ellinghaus D¹²) as well as information for the exact same variants or variants in high-LD ($R^2 > 0.9$) from the publicly available RICOPILI summary statistics (IBD_UC_1KG_oct13). The specific variant is characterized by its rs-id, the chromosomal position in

GRCh38 (hg38) and GRCh37 (hg19) as well as the given alleles. From the association analysis the P-value (p.value) the OR with its 95% confidence interval (CI_L95 and CI_U95) as well as the beta and standard error (SE) are given. Further, the MAPPED_TRAIT is shown, which is always ulcerative colitis for our own data and the dataset extracted from RICOPILI but varies for data from the GWAS catalog as also variants including IBD are listed in case no association with UC could be identified for this locus in the database but another IBD related trait. Further the allele frequencies separated by patients (AF.Cases) and controls (AF.Controls) is given. For the RICOPILI data the frequencies are as given by Hapmap. For imputed variants, the imputation info is given. The R2 is related to imputed genotyping lead variant, the corresponding dataset is also given in the dataset_Id column. The mapped gene and the related impact on the protein structure (change) are listed as well as gtx associated genes associated with the variant.

[See separate xlsx-file]

Supplementary Table 4: The association results with the HLA imputed data. In the “type” column the type of the analyzed variant is given as one of the following: 1-field HLA, 2-field HLA, nuc, or prot. The HLA gene name (locus) and the exact description of the variant (name) together with the chromosome (CHR) and position in the human reference genome GRCh38 (hg38) and GRCh37 (hg19) and the position within the protein sequence in case of nucleotides (nuc) and amino acids (prot) specify the exact variant. The alleles A1 and A2 are either the one letter nucleotide or amino acid code or in the case of polymorphic variants and HLA alleles noted as present-absent (P and A). From the association analysis the P-value (p.value), the OR with its 95% confidence interval (CI_L95 and CI_U95), as well as the standard error (SE) are given. Further the minor allele frequency (MAF) as well as the allele frequency separated by cases (AF.Cases) and controls (AF.Controls) is included. Additionally, for the HLA alleles the posterior probability is listed as a reliability score of the imputation.

[See separate xlsx-file]

Supplementary Table 5: The significant associated peptides from the PepWAS analysis (**sheet: peptides**) and the information condensed on the transcript (**sheet: transcripts**) and gene level (**sheet: genes**). For each peptide the association statistics are given as effect size (effect_size) and p-value (p) further the corresponding transcripts (ENSTs) with the position in GRCh38, the ensemble gene ids (ENSG) and the hgnc symbol is given. It is noted whether the peptide is present in the reference proteome (mutations) and in how many samples (n). The sample numbers with a specific peptide are also listed separated by cases (n_nase) and controls (n_controls). Those values are also given as frequencies (f, f_case and f_control). Also, the OR of the frequencies (OR_freq) is included and the p-value of the fisher test on the frequencies. The mutations related to the single peptides are listed: First the mutations that are needed to generate the sequence (nucchanges and protchanges), further the different amino acid positions in different transcripts are listed and then all mutations, that would change the sequence of the peptide, are listed. The column HLA lists the HLA alleles predicted to bind the peptide. The number (n_immunopeptidome_blood) and sequences (immunopeptidome_blood) of identified peptides in the 25 immunopeptidomes published in ElAbd et al.¹³ are presented. On the transcript level (**sheet: transcripts**) for each ensemble transcript (ENST) the uniprot gene id and the ensemble protein id together with the length of the amino acid sequence (lengthAA) are noted. The number of peptides per transcript are given (n_hits). As the peptides are generated by a sliding window approach and peptides binding HLA class II are longer than the binding pocket, often neighboring peptides are predicted as similar good peptides and a single missense mutation might have only a small impact on the binding affinity, therefore also the number of peptides where less than 9 amino acids are in the same order are given (n_no9AAoverlap). Further, the number of PepWAS hits is separated by those with a mutation (n_hits_mut) and those present in the reference (n_hits_ref). Additional numbers of related mutations (n_relevant_mut) are given as mutations necessary to form a PepWAS hit (necessary_mut), mutations that are changing the present peptide but both peptides are PepWAS hits (possible_mut), and mutations that modify a peptide in a way that it is not predicted as PepWAS hit anymore (forbidden_mut). On the gene level (**sheet: genes**) the different transcripts are summarized, with the majority of the previously described attributes, and whether all peptides annotated to one gene are expressed within one transcript (AllHitsInOneENST). Further information about the expression of the genes is given as reported in Taman et al.¹⁴ and in Linggi et al.¹⁵. Further if the confidence set of associated variants includes any GTEx¹⁶ variants the effect is given in comparison to the risk variants (risk_variant_lead_to_expression) as “decreased” or “increased” expression. The immunopeptidome data are given for the whole genes independent of the location of the PepWAS hits.

[See separate xlsx-file]

References

1. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
2. Pruim, R. J. *et al.* LocusZoom: Regional visualization of genome-wide association scan results. *Bioinformatics* **27**, 2336–2337 (2011).
3. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).
4. De Lange, K. M. *et al.* Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* **49**, 256–261 (2017).
5. Silverberg, M. S. *et al.* Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study. *Nat. Genet.* **41**, 216–220 (2009).
6. McGovern, D. P. B. *et al.* Genome-wide association identifies multiple ulcerative colitis susceptibility loci. *Nat. Genet.* **42**, 332–337 (2010).
7. Jostins, L., Ripke, S., Weersma, R. K., Duerr, R. H. & Dermot, P. Host-microbe interactions have shaped the genetic architecture of Inflammatory Bowel Disease. **491**, 119–124 (2012).
8. Barrett, J. C. *et al.* Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. *Nat. Genet.* **41**, 1330–1334 (2009).
9. Anderson, C. A. *et al.* Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat. Genet.* **43**, 246–252 (2011).
10. Asano, K. *et al.* A genome-wide association study identifies three new susceptibility loci for ulcerative colitis in the Japanese population. *Nat. Genet.* **41**, 1325–1329 (2009).
11. Okamoto, D. *et al.* Genetic analysis of ulcerative colitis in Japanese individuals using population-specific SNP array. *Inflamm. Bowel Dis.* **26**, 1177–1187 (2020).
12. Ellinghaus, D. *et al.* Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. *Nat. Genet.* **48**, 510–518 (2016).
13. Elabd, H. *et al.* Predicting Peptide HLA-II Presentation Using Immunopeptidomics , Transcriptomics and Deep Multimodal Learning. *bioRxiv* (2022) doi:<https://doi.org/10.1101/2022.09.20.508681>.
14. Taman, H. *et al.* Transcriptomic landscape of treatment-naïve ulcerative colitis. *J. Crohn's Colitis* **12**, 327–336 (2018).
15. Linggi, B. *et al.* Meta-analysis of gene expression disease signatures in colonic biopsy tissue from patients with ulcerative colitis. *Sci. Rep.* **11**, 1–12 (2021).
16. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).